# Effekte von Testteilnahmemotivation auf Testleistung im Kontext von Large-Scale-Assessments

Dissertation

zur Erlangung des akademischen Grads

Dr. phil.

im Fach Erziehungswissenschaften

eingereicht am 27. Februar 2015

verteidigt am 11. Mai 2015

an der Kultur-, Sozial- und Bildungswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin

von

Christiane Penk, M. A.

Präsident der Humboldt-Universität zu Berlin

Prof. Dr. Jan-Hendrik Olbertz

Dekanin der Kultur-, Sozial- und Bildungswissenschaftlichen Fakultät

Prof. Dr. Julia von Blumenthal

Begutachtung durch

1. Prof. Dr. Petra Stanat
2. Prof. Dr. Olaf Köller

# Inhalt

## Zusammenfassung

Die vorliegende Arbeit untersucht die Testteilnahmemotivation von Schülerinnen und Schülern in großangelegten Schulleistungsstudien. Bisherige Forschung in diesem Bereich basiert zwar theoretisch auf dem Erwartung-Wert-Modell der Leistungsmotivation, entwickelte jedoch kein an die Besonderheiten der Testteilnahmemotivation angepasstes Modell. Daher wurde in dieser Arbeit ein theoretisches Erwartung-Wert-Anstrengung-Modell der Testteilnahmemotivation herausgearbeitet, das in drei empirischen Studien überprüft wurde. In diesem Modell wird Anstrengungsbereitschaft explizit als Ergebnis von Erwartung und Wert betrachtet; Anstrengung, Erwartung und Wert werden mit der Testleistung in Verbindung gebracht. In der Arbeit war vor allem das bisher unerforschte komplexe Beziehungsgefüge zwischen Erfolgserwartungen, dem wahrgenommene Wert des Tests, Anstrengungsbereitschaft und Testleistung von Interesse.

Die drei Studien basieren auf zwei realistischen Testsituationen. Datengrundlage der Studie I bildete die erste PISA-Erhebung aus dem Jahr 2000, in der die Testteilnahmemotivation durch Fragen zur Anstrengungsbereitschaft und zum wahrgenommenen Wert des Tests erhoben wurde. In Studie II und III gaben die Schülerinnen und Schüler, die an der Ländervergleichsstudie im Jahr 2012 teilnahmen, Einschätzungen zu ihren Erfolgserwartungen, dem wahrgenommenen Wert des Tests und ihrer Anstrengungsbereitschaft ab.

Insgesamt zeigen die Ergebnisse, dass Testteilnahmemotivation zur Erklärung individueller Unterschiede in der Testleistung beiträgt (Studie I), auch wenn diverse Hintergrundinformationen der Teilnehmenden berücksichtigt werden (Studie III). Die theoretisch angenommenen Beziehungen im Erwartung-Wert-Anstrengung-Modell wurden fast vollständig bestätigt: Vor allem der wahrgenommene Wert, aber auch die Erwartungen sagten die berichtete Anstrengungsbereitschaft der Teilnehmenden vorher; die Erfolgserwartungen und die Anstrengungsbereitschaft wiesen einen Zusammenhang mit der Testleistung auf (Studie II). Im Verlauf eines Leistungstests berichteten die Teilnehmenden im Durchschnitt eine Abnahme in der Anstrengung und dem Wert sowie einen stabilen Verlauf ihrer Erfolgserwartungen. Zur Erklärung der Testleistung trug neben den vor dem Test berichteten Erfolgserwartungen und Anstrengungsbereitschaft auch die Veränderung in den Erfolgserwartungen bei (Studie III). Für eine hohe Testleistung ist es wichtig, dass die Teilnehmenden den Test motiviert beginnen und während des Tests selbstsicher bezüglich ihrer Erfolgserwartungen bleiben. Zusammenfassend sollten alle drei Komponenten erfasst werden, um Testteilnahmemotivation vollständig zu modellieren.

## Abstract

The thesis investigates effects of test-taking motivation on test performance in low-stakes assessments. Low-stakes tests have no consequences for test-takers so that unmotivated behavior can threaten a valid interpretation of the test results. Previous research was based on expectancy-value theory of achievement motivation. However, this model does not take into account the specifics of test-taking motivation. For that reason, an expectancy-value-effort model of test-taking motivation was developed and tested in three empirical studies. In this expectancy-value-effort model, effort is modeled as the outcome of both expectancy and value; in turn, effort, expectancy, and value are related to test performance. The three studies investigated the complex relationship between expectancy for success, perceived value of the test, test-taking effort, and test performance to broaden the existing theory.

The three studies were based on two realistic testing situations. The database of study I is the first PISA study. Test-taking motivation was assessed with questions about effort and the perceived value of the test. Study II and III are premised on the national assessment study in the year 2012. The students reported their expectancy for success, their perceived value of the test, and their test-taking effort.

Overall, the results showed that test-taking motivation explained test performance (study I) although controlling for various students' background characteristics (e.g., socio-economic background, study III). We found support for nearly all of the theoretically assumed relationships in the expectancy-value-effort model: Expectancy for success and perceived value of the test explained test-taking effort; expectancy for success and test-taking effort had the most pronounced effects on test performance (study II). The students reported, on average, a stable course of expectancy for success over the testing session; perceived importance of the test and test-taking effort slightly decreased within the testing session. The initial expectancy for success and the initial test-taking effort as well as change in expectancy for success explained students' test performance. Above all, it is crucial that students begin the test with a high level of test-taking motivation and remain confident about a successful test completion to the end of the testing session. In sum, the results point out that a comprehensive model of test-taking motivation should include all three components of expectancy, value, and test-taking effort.

# 1

# Problemstellung

# 1 Problemstellung

Die Qualität von Bildungssystemen ist für die Bildungspolitik weltweit von zentraler Bedeutung. Schon Ende der 1950er-Jahre begann die *International Association for the Evaluation of Educational Achievement* (IEA) mit der Erhebung und Auswertung von Schülerleistungsdaten. Heute stellt die Organisation für wirtschaftliche Zusammenarbeit und Entwicklung (*Organisation for Economic Co-operation and Development*, OECD) eine Vielzahl von Studien im Bildungsbereich bereit. Die Ergebnisse dieser Untersuchung fungieren unter anderem als Indikatoren, um die Kompetenzen der Schülerinnen und Schüler und damit das Leistungspotential der Bildungssysteme der OECD-Mitgliedsstaaten zu vergleichen (Böhm-Kasper & Weishaupt, 2008). In den letzten zwei Jahrzehnten gewannen groß angelegte Schulleistungsstudien (sogenannte *Large-Scale-Assessments*) auch für die Beurteilung des deutschen Bildungssystems an Bedeutung. Deutschland beteiligt sich seit Ende der 1990er-Jahre regelmäßig an internationalen und nationalen Schulleistungsstudien, wie beispielsweise dem *Programme for International Student Assessment* (PISA), der *Trends in Mathematics and Science Study* (TIMSS) oder der Internationalen Grundschul-Lese-Untersuchung (IGLU). Nach dem sogenannten „PISA-Schock“ im Jahr 2001 beschloss die Kultusministerkonferenz in Deutschland die Einführung national verbindlicher Bildungsstandards als Reaktion auf die unterdurchschnittlichen Ergebnisse der Schülerinnen und Schüler in PISA beziehungsweise auf die mittelmäßige Leistung in TIMSS. Diese Standards definieren den erwünschten Kompetenzstand der Lernenden in Deutsch und Mathematik zum Ende der Grundschule und in Mathematik sowie den naturwissenschaftlichen Fächern (Biologie, Chemie, Physik) und den Sprachen (Deutsch, Englisch, Französisch) zum Ende der Sekundarstufe I (z. B. KMK, 2004, 2005a-d). Zur Überprüfung der Bildungsstandards werden seit 2009 Ländervergleiche in der vierten und neunten Jahrgangsstufe durchgeführt.

Ein wesentliches Ziel großer internationaler Leistungsstudien ist die Bereitstellung von Informationen über die Stärken und Schwächen der Schülerinnen und Schüler des eigenen Bildungssystems im Kontrast zu den Stärken und Schwächen der Schülerinnen und Schüler anderer Bildungssysteme. Dadurch soll steuerungsrelevantes Wissen für die Bildungspolitik zur Verfügung gestellt werden, um politisch-administrative Entscheidungen zur Verbesserung des nationalen Bildungssystems zu unterstützen. Anhand der Ergebnisse der verschiedenen Leistungsstudien wird insgesamt die Qualität des Bildungs-

systems evaluiert und es werden beispielsweise Rankings erstellt, die in bildungspolitische Entscheidungsfindungsprozesse einfließen können (Schwippert & Goy, 2008; Stanat & Lüdtke, 2013). Eine Voraussetzung für eine valide Interpretation der Ergebnisse und die Formulierung gültiger Aussagen über die Kompetenzen der Schülerinnen und Schüler ist, dass sie während der Testsitzung ihr Bestes geben und demzufolge eine maximale Performanz zeigen. Jedoch haben diese Art von Tests keine positiven oder negativen Konsequenzen für die Testteilnehmenden, unabhängig davon wie gut oder schlecht diese abschneiden. Es handelt sich bei diesen groß angelegten Schulleistungsstudien um sogenannte *Low-Stakes-Assessments*, bei denen die Testteilnehmenden weder eine Note für ihre Leistung noch individuelles Feedback bekommen. Die Vorstellung, dass die Schülerinnen und Schüler den Test unter bestmöglicher Anstrengung bearbeiten, stellt ein Idealbild einer Testteilnehmenden beziehungsweise eines Testteilnehmenden dar. Tatsächlich kann allerdings nicht davon ausgegangen werden, dass die Kinder und Jugendlichen an die Testaufgaben motiviert herangehen, selbst wenn sie die notwendigen Fähigkeiten für eine erfolgreiche Bearbeitung der Testaufgaben besitzen. In diesem Fall ist die Validität der Interpretation und Nutzung der Testergebnisse gefährdet (Asseburg, 2011; Eklöf, 2007, 2008, 2010a, 2010b; Thelk, Sundre, Horst & Finney, 2009).

Um die Bedeutung einer Gefährdung der Validität nachvollziehen zu können, soll zunächst geklärt werden, was in diesem Kontext Validität bedeutet. Nach dem weit verbreiteten Validitätskonzept von Messick (1989) ist Validität „(…) an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores or other modes of assessment" (S. 13, Hervorhebungen im Original). Dadurch soll deutlich werden, dass nicht der Test valide sein kann, sondern nur die spezielle Verwendung und Interpretation der Testergebnisse (Linn, 2010). Nach Messick (1989) sind die größten Gefahren, die eine valide Interpretation und Nutzung der Testergebnisse einschränken, die Unterrepräsentation des Konstruktes und konstrukt-irrelevante Varianz. Wenn der empirische Test zu eng gefasst ist und daher nicht alle relevanten Aspekte des theoretischen Konstrukts erfasst werden, liegt eine Unterrepräsentation des Konstrukts vor. Konstrukt-irrelevante Varianz ist vorhanden, wenn mit dem konstruierten Test Inhalte erhoben werden, die nicht zum Kern des Konstrukts gehören (Messick, 1989). Dazu gehört zum Beispiel die Testteilnahmemotivation bei der Bearbeitung von Low-Stakes-

Assessments. Es besteht demnach beispielsweise die Gefahr, dass ein Mathematiktest neben der Mathematikkompetenz auch Testteilnahmemotivation misst.

So weist die aktuellste Version der *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 2014) darauf hin, dass unter anderem Testteilnahmemotivation bedacht werden sollte, wenn die Ergebnisse von Low-Stakes-Assessments berichtet und interpretiert werden (Standards 1.10, 3.18, 4.10, 9.13, 10.12, 13.6 und 13.9). „When there is a suspicion that the test might not have been taken seriously, the motivation of test takers may be explored by collecting additional information where feasible, using observation or interview methods. Issues of inappropriate preparation or unmotivated performance raise questions about the validity of interpretations of test results. In every case, it is important to consider the potential impact on the test taker of the testing process itself, including test administration and reporting practices" (AERA et al., 2014, S. 207). Gerade bei Low-Stakes-Assessments liegt der Verdacht nahe, dass nicht alle Teilnehmenden den zu bearbeitenden Test ernst nehmen und die Aufgaben nach bestem Wissen und Gewissen bearbeiten. Auch die *International Test Commission* macht darauf aufmerksam, mögliche Einflüsse zu bedenken, die Testergebnisse senken oder erhöhen können, wie die Testteilnahmemotivation (International Test Commission, 2001).

Diese Arbeit befasst sich mit der Untersuchung von Effekten der Testteilnahmemotivation auf die Testleistung in Large-Scale-Assessments als potentielle Gefährdung der validen Interpretation und Nutzung der Testergebnisse. Ohne Kenntnisse darüber, ob die gewonnenen Testergebnisse unverfälschte Messungen der Fähigkeiten der Testteilnehmenden darstellen, ist die Evaluation des Bildungssystems und die Ableitung geeigneter bildungspolitischer Handlungsoptionen nicht zuverlässig möglich. Die hier skizzierte Problemstellung bildet den praktischen Ausgangspunkt der vorliegenden Arbeit, der die Auseinandersetzung mit dem Thema Testteilnahmemotivation auf Basis der Erwartung-Wert-Theorie angeregt hat. Die theoretische Grundlage der Arbeit wird im folgenden Abschnitt beschrieben.

# 2

# Theoretischer Rahmen

# 2 Theoretischer Rahmen

In der Motivationsforschung wird eine Reihe von Begriffen unterschieden, die zum Verständnis und zur Einordnung der Testteilnahmemotivation zunächst im Abschnitt 2.1 genauer beschrieben werden. Zuerst werden Motivation und Leistungsmotivation definiert und konzeptionell voneinander abgegrenzt, um anschließend das Grundmodell der „klassischen“ Motivationspsychologie vorzustellen. Dieses Modell bildet die Grundlage zum Verständnis leistungsmotivierten Verhaltens in der Motivationspsychologie. Anschießend werden verschiedene Theorien domänenspezifischer Leistungsmotivation skizziert (Abschnitt 2.2), deren Überschneidungen mit der Erwartung-Wert-Theorie (Abschnitt 2.3) später aufgegriffen werden. Nachdem das Risikowahl-Modell als Vorläufer des Erwartung-Wert-Modells beschrieben wurde, wird auf das Erwartung-Wert-Modell der Leistungsmotivation eingegangen. In Abschnitt 2.4 folgt die Definition und die theoretische Verortung der Testteilnahmemotivation sowie die Beschreibung des Forschungsstandes, um letztlich ein an die Besonderheiten der Testteilnahmemotivation angepasstes Erwartung-Wert-Anstrengung-Modell zu entwerfen.

Abbildung 2.1 dient der übergreifenden Illustration der im Theorieteil beschrieben Konstrukte und Modelle sowie deren Zusammenhänge untereinander und theoretische Einordnung. Ausgehend von der allgemeinen Motivationsdefinition und der Leistungsmotivation sowie dem Erwartung-Wert-Modell der Leistungsmotivation wird zum Erwartung-Wert-Anstrengung-Modell der Testteilnahmemotivation geleitet, das den theoretischen Ausgangspunkt der Fragestellungen dieser Arbeit bildet. In der Abbildung wird bereits deutlich, dass die in Abschnitt 2.2 erläuterten Konstrukte (Selbstwirksamkeit, Selbstkonzept, Interesse und Ziele) teilweise sowohl der situationsspezifischen als auch der domänenspezifischen Leistungsmotivation zugeordnet werden können. Außerdem wird verbildlicht, dass die verschiedenen Theorien nicht lose nebeneinander, sondern miteinander in Beziehung stehen und daher auch Überschneidungen aufweisen. Diese Überschneidungen und auch Unterschiede werden im folgenden Theorieteil erläutert.

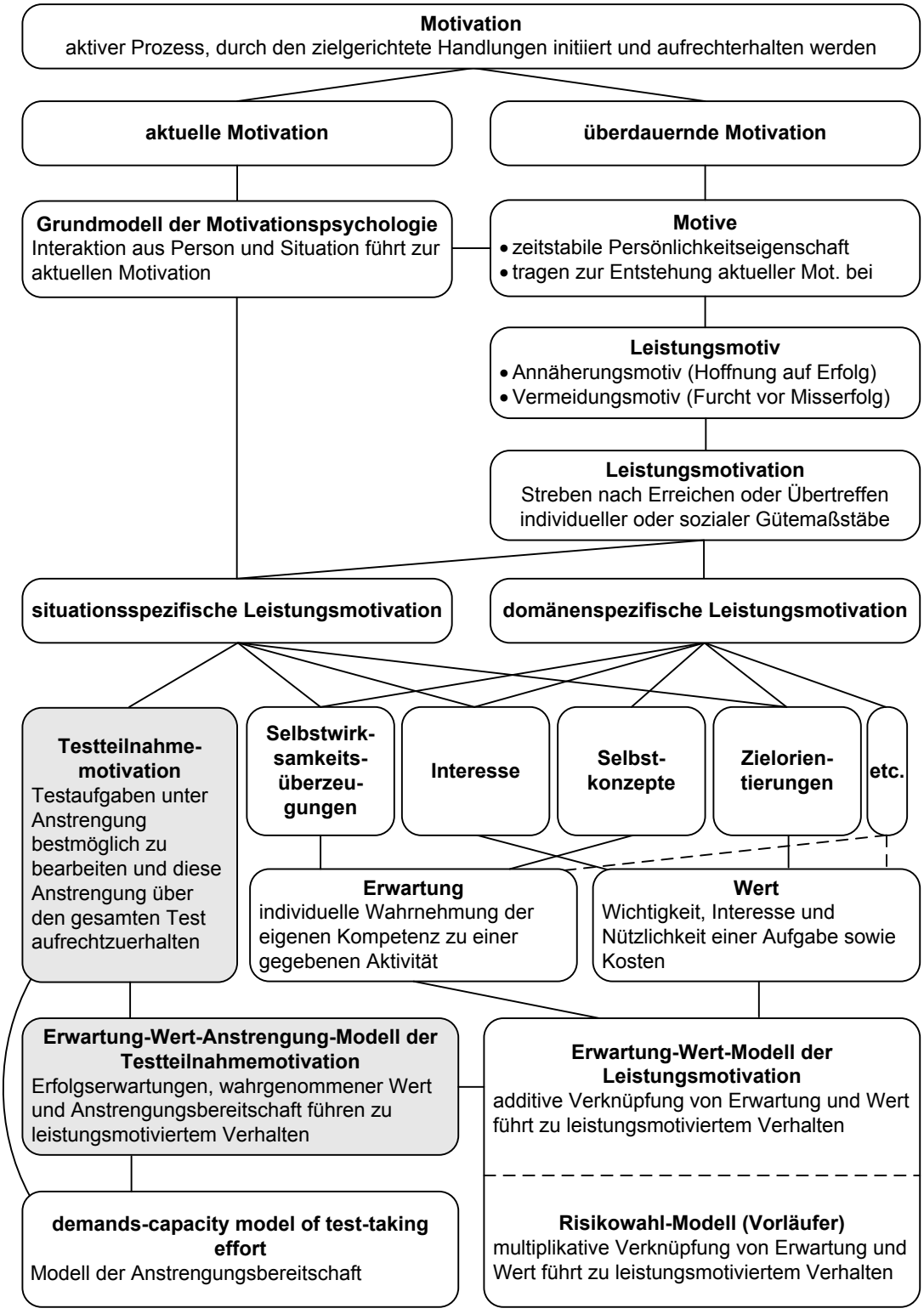



*Abbildung 2.1*. Übersicht der im Theorieteil beschriebenen Konstrukte und Modelle sowie deren theoretische Zusammenhänge.

## 2.1 Begriffsbestimmung und Überlegungen zu leistungsmotiviertem Verhalten

### 2.1.1 Motivation

Der Wortursprung des Motivationsbegriffs geht zurück auf das lateinische Wort *movere*, das bewegen oder auch anregen bedeutet. Diese dynamische Komponente ist ebenfalls in der allgemeinen Definition von Rheinberg (2008, S. 15) erkennbar, nach der Motivation als „aktivierende Ausrichtung des momentanen Lebensvollzugs auf einen positiv bewerteten Zielzustand“ ausgefasst wird und verschiedene Teilprozesse beinhaltet. Dabei sollen mithilfe der Motivationspsychologie die Beweggründe für motiviertes Verhalten sowie das Zusammenwirken der Teilprozesse erklärt werden, die die Richtung, Ausdauer und Intensität der Verhaltensweisen bestimmen (Rheinberg, 2008). Schiefele (2009) differenziert grundlegend zwischen aktueller und überdauernder Motivation und ordnet insbesondere der überdauernden Motivation eine Vielzahl von Konstrukten zu. Eine Konzeptualisierung überdauernde Motivation bezieht sich auf den Begriff des Motivs. Motive werden als überdauerndes Personenmerkmal definiert, die darauf Einfluss nehmen, wie eine Person bestimmte Situationen wahrnimmt und beurteilt (Rheinberg, 2008; Schiefele, 2009). Angeregt durch Merkmale der konkreten Situation tragen Motive auch zur Entstehung aktueller Motivation bei, wie in Abbildung 2.1 dargestellt. Dieser Prozess steht beispielsweise in Erwartung-Wert-Modellen im Fokus (s. Abschnitt 2.3.2). Neben der Forschung zu Motiven gibt es eine Vielzahl anderer Forschungsstränge, die der überdauernden Motivation zugeordnet werden können, wie zum Beispiel Zielorientierungen und Interesse oder domänenspezifische Selbstkonzepte, die in Abschnitt 2.2 umrissen werden (Schiefele, 2009).

### 2.1.2 Leistungsmotivation

Da hier Motivation im Schulkontext, konkret die Motivation für die Teilnahme an Schulleistungsstudien untersucht wird, muss die allgemeine Motivationsdefinition zunächst für den Leistungskontext präzisiert werden, denn nicht jede Anstrengung ist per se leistungsmotiviert. Leistungsmotivation ist durch das individuelle Streben gekennzeichnet, individuell gesetzte Gütemaßstäbe erreichen beziehungsweise übertreffen zu wollen (Heckhausen & Heckhausen, 2006). In der Auseinandersetzung mit diesem Gütemaßstab wird die eigene Aktivität bewertet. Dabei genügt als Anregung für die Zielerreichung das Gefühl, eine Tätigkeit eigenständig bewältigt zu haben (Rheinberg, 2008). An dieser Stelle

soll vorweggenommen werden, dass die Erreichung des individuell gesetzten Gütemaßstabs im Kontext der Testteilnahmemotivation nicht überprüft werden kann, da die Schülerinnen und Schüler bei der Bearbeitung von Low-Stakes-Tests keine Rückmeldung über ihr Ergebnis erhalten.

Leistungsmotivation kann weiter unterteilt werden in domänenspezifische und situationsspezifische Leistungsmotivation, wobei umfangreiche Forschung zur domänenspezifischen Form zu finden ist (Eklöf, 2010a). Die Unterscheidung zwischen domänenspezifischer Motivation und situationsspezifischer Motivation korrespondiert mit der in Abschnitt 2.1.1 genannten Differenzierung zwischen überdauernder (d. h. domänenspezifischer) und aktueller (d. h. situationsspezifischer) Motivation (Schiefele, 2009). Der domänenspezifischen Form ist beispielsweise die Erfassung der Leistungsmotivation im Fach Mathematik zuzuordnen; im Gegensatz dazu gehört die Motivation, in einer gegebenen Situation gut abzuschneiden, zur situationsspezifischen Form.

### 2.1.3 Personen- und Situationsmerkmale

Aktuelle Motivation entsteht erst durch die situative Anregung eines Motivs. Das Grundmodell der „klassischen“ Motivationspsychologie (Rheinberg, 2008) trägt zur Erklärung leistungsmotivierten Verhaltens bei und wird daher in diesem Abschnitt vorgestellt. Die klassische Motivationstheorie nimmt an, dass sowohl die Person als auch die Situation Einfluss auf die aktuelle Motivation haben und so leistungsmotiviertes Verhalten bedingen (Heckhausen & Heckhausen, 2006; Rheinberg, 2008). Dabei beziehen sich die Personenmerkmale auf Motive (hier konkret auf das Leistungsmotiv), die relativ zeitstabile, überdauernde Persönlichkeitseigenschaften darstellen, die unabhängig von der Situation ähnlich auf das Leistungsverhalten einer Person einwirken. Jedoch verhalten sich Personen in unterschiedlichen Situationen nicht immer gleich, weshalb die Berücksichtigung situativer Merkmale von Bedeutung ist. Situationsmerkmale sind situationsspezifische intrinsische und extrinsische Anreize mit Aufforderungscharakter (Heckhausen & Heckhausen, 2006), die an verschiedene Aspekte von Handlungsfolgen geknüpft sind. Dabei können auch Testmerkmale Einfluss auf die Motivation ausüben, wie beispielsweise die Aufgabenschwierigkeit (Asseburg, 2011).

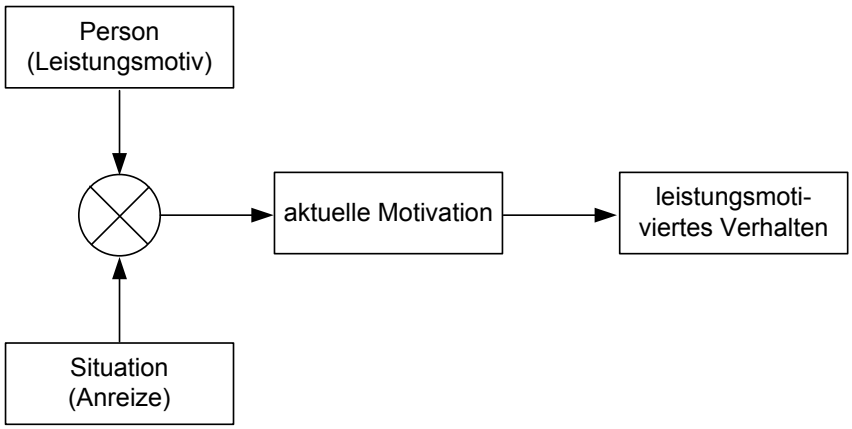



*Abbildung 2.2*. Grundmodell der „klassischen“ Motivationspsychologie unter Berücksichtigung des Leistungsmotivs (in Anlehnung an Rheinberg, 2008, S. 70).

„Typisch für die klassische Motivationspsychologie ist also eine Trennung von Motiv als überdauerndem Personenmerkmal und der je aktuellen Motivation, die aus der Wechselbeziehung zwischen Situation und Motiv resultiert“ (Rheinberg, 2008, S. 70). Nach dem Modell mündet das Leistungsmotiv nur dann in motiviertes Verhalten, wenn sich die Person auch in einer motivierenden Situation befindet. Dies wird in Abbildung 2.2 durch die Interaktion verdeutlicht. Erst durch diese Interaktion von Leistungsmotiv auf der Personenseite und den potentiellen Anreizen auf der Situationsseite entsteht die aktuelle Motivation, die sich in entsprechendem leistungsmotiviertem Verhalten widerspiegelt.

## 2.2 Theorien der Leistungsmotivation

Die Vielzahl an existierenden Konzeptualisierungen zu leistungsmotiviertem Verhalten wurde von Pintrich, Marx und Boyle (1993) in zwei sehr breite Kategorien eingeordnet. Zur ersten Kategorie gehören Konstrukte, die sich auf Fähigkeits- oder Kompetenzüberzeugungen beziehen, verschiedene Tätigkeiten ausführen zu können. Diese Überzeugungen beziehen sich allgemein auf die *Erwartungen* von Personen. Die zweite Kategorie bündelt Konstrukte zu den Gründen für die Ausführung von Tätigkeiten und verweist damit auf den *Wert*, der einer bestimmten Handlung zugeschrieben wird. *Erwartung-Wert-Modelle* wiederum integrieren sowohl Konstrukte zur Erwartung als auch Konstrukte zum Wert. Bevor die Erwartung-Wert-Theorie in Abschnitt 2.3 vorgestellt wird, werden zwei Theorien beschrieben, die jeweils nur auf eine Komponente, also entweder nur auf die Erwartung oder nur auf den Wert fokussieren. Außerdem wird kurz der empirisch gefundene Zusammenhang zwischen den vorgestellten Konstrukten und Leistung aufgezeigt. Im nächsten Abschnitt werden dann Überschneidungen und Divergenzen dieser Theorien zum

Erwartung-Wert-Modell herausgearbeitet. Dabei erhebt dieses Unterkapitel keinen Anspruch auf Vollständigkeit, sondern soll vielmehr exemplarisch aufzeigen, welche Fülle an Theorien zur Erklärung leistungsmotivierten Verhaltens existiert.

### 2.2.1 Kompetenzüberzeugungen

Viele Motivationstheorien beschreiben die Vorstellungen und Überzeugungen von Individuen bezüglich der Wahrnehmung ihrer eigenen Kompetenzen, Wirksamkeiten und Erfolgserwartungen. Damit wird leistungsmotiviertes Verhalten auf die Einschätzung der eigenen Kompetenz zurückgeführt und nicht auf die „tatsächliche“ Fähigkeit der Individuen. Beispielsweise beschreibt Bandura (1997) mit seinem Selbstwirksamkeitskonstrukt die Rolle der wahrgenommenen Fähigkeit für den Erfolg eigener Handlungen. Dabei wird Selbstwirksamkeit definiert als die individuelle Wahrnehmung der eigenen Fähigkeit, um Probleme zu lösen oder eine Aufgabe zu bewältigen (Eccles & Wigfield, 2002). Bandura (1997) unterscheidet in diesem Kontext zwei Arten von Erwartungsüberzeugungen: Ergebniserwartungen und Wirksamkeitserwartungen. Ergebniserwartungen beziehen sich auf Überzeugungen, dass ein spezifisches Verhalten zu einem bestimmten positiven oder negativen Ergebnis führt. Im Gegensatz dazu beschreiben Wirksamkeitserwartungen die Überzeugung einer Person, dieses Verhalten auch ausführen zu können (Bandura, 1977). Insgesamt beinhalten Selbstwirksamkeitsüberzeugungen deskriptive, zielbezogene, relativ kontextspezifische und zukunftsorientierte Beurteilungen der eigenen Kompetenz, die durch ihre Situations- und Aufgabenabhängigkeit formbar sind (Bong & Skaalvik, 2003). Studien haben gezeigt, dass hohe Selbstwirksamkeitsüberzeugungen einen positiven Effekt auf Leistungen haben (Pajares & Graham, 1999; Pajares & Kranzler, 1995). Für eine hohe Anstrengungsbereitschaft, Ausdauer sowie Erfolgswahrscheinlichkeit ist eine positive Wahrnehmung der Wirksamkeit der eigenen Handlungen bezüglich der zu bewältigenden Aufgabe von besonderer Relevanz (Schunk & Pajares, 2005).

Ein verwandtes Konstrukt im Kontext von Kompetenzüberzeugungen ist das akademische Selbstkonzept (Köller, Trautwein, Lüdke & Baumert, 2006; Marsh, 1990), das als individuelle Einschätzung der Fähigkeit einer Person in einem bestimmten Fach definiert werden kann (Stiensmeier-Pelster & Schöne, 2008). Im Gegensatz zu Selbstwirksamkeitsüberzeugungen sind Selbstkonzeptüberzeugungen evaluative, durch die Vergleiche der eigenen Fähigkeit mit der Fähigkeit Anderer eher normative, typischerweise aggregierte, hierarchisch strukturierte und vergangenheitsorientierte Selbstwahrnehmungen, die relativ stabil sind. Selbstwirksamkeit kann als ein aktiver Vorläufer der Selbstkonzeptentwicklung

verstanden werden (Bong & Skaalvik, 2003). Individuen mit hohen Selbstwirksamkeitsüberzeugungen neigen dazu, günstige akademische Selbstkonzepte aufzuweisen. Neben Selbstwirksamkeitsüberzeugungen leisten vor allem soziale, dimensionale, temporale und kriteriale Vergleichsinformationen einen Beitrag zur Bildung des individuellen akademischen Selbstkonzepts (Möller & Trautwein, 2009; Schunk & Pajares, 2005). Bezüglich des Zusammenhangs zwischen dem domänenspezifischen Selbstkonzept und der Leistung zeigen Studien, dass beispielsweise das mathematisches Selbstkonzept positiv mit der Mathematiknote sowie mit Ergebnissen aus standardisierten Mathematiktests assoziiert ist (Jansen, Schroeders & Stanat, 2013; Marsh & Yeung, 1998).

### 2.2.2 Interesse und Ziele

Motivationstheorien mit dem Schwerpunkt auf Kompetenzüberzeugungen liefern aussagekräftige Erklärungen für individuelle Leistungen bei verschiedenen Aufgaben. Jedoch betrachten diese Theorien nicht systematisch die Gründe, warum sich Personen bei unterschiedlichen Aufgaben engagieren unabhängig davon, ob sie glauben, die Aufgabe bewältigen zu können. Dazu gehören beispielsweise Theorien zum Interesse, das als interaktive Beziehung zwischen einem Individuum und bestimmten Aspekten aus dessen Umwelt (z. B. einem Objekt oder Ereignis) definiert werden kann (Hidi & Harackiewicz, 2000). Dabei wird zwischen individuellem und situativem Interesse unterschieden. Individuelles Interesse ist eine relativ stabile motivationale Orientierung oder persönliche Disposition, die sich über die Zeit hinweg und in Relation mit einem bestimmten Gegenstandsbereich entwickelt. Das individuelle Interesse kann weiter unterteilt werden in gefühlsbezogene Valenzüberzeugungen, die auftreten, wenn der Bereich mit positiven Emotionen verknüpft ist, und in wertbezogene Valenzüberzeugungen, die präsent sind, wenn die Person dem Bereich persönliche Wichtigkeit zuschreibt. Im Kontrast dazu ist situatives Interesse ein emotionaler Zustand, der durch bestimmte Merkmale einer Aufgabe oder Tätigkeit hervorgerufen wird und durch Neugier, Aufmerksamkeit oder Begeisterung gekennzeichnet ist (Hidi & Harackiewicz, 2000; Schiefele, 1999). Die meisten Forschungsarbeiten zu situativem Interesse beschäftigten sich mit Merkmalen von Textaufgaben, die bei den Individuen situatives Interesse hervorrufen, sowie mit dem Leseverstehen (Schiefele, 2009). Insgesamt belegen Studien, dass Interesse nur schwach mit Leistung zusammenhängt, dafür aber mit Wahlentscheidungen, wie beispielsweise mit der Entscheidung für einen Leistungskurs (Harackiewicz, Durik, Barron, Linnenbrink-Garcia & Tauer,

2008; Köller, Baumert & Schnabel, 2001; Marsh, Trautwein, Lüdtke, Köller & Baumert, 2005).

Zielorientierungen beziehen sich auf die Herangehensweisen, die Individuen zum Lernen einnehmen, und werden als „mental repräsentierte Zielvorstellungen konzeptualisiert“ (Schiefele, 2009, S. 154). Sie bilden eines der klarsten Konstrukte bezüglich individueller Zwecke und Ziele zur Kompetenzentwicklung (Wigfield & Cambria, 2010). Zieleorientierungen können als explizite und kognitiv repräsentierte Dispositionen definiert werden, die allerdings auch situativ ausgelöst werden können (Schiefele, 2009). Bedeutend für Forschung in diesem Bereich sind die etwa zeitgleich entwickelten Theorien von Nicholls (1984) und Dweck (1986), die zwischen Lernzielen (*ego orientation* bzw. *learning/mastery goals*) und Leistungszielen (*task orientation* bzw. *performance goals*) unterscheiden. Eine an Lernzielen orientierte Person strebt danach, die eigenen Fähigkeiten zu verbessern und neue Dinge zu erlernen. Im Gegensatz dazu bezweckt eine an Leistungszielen orientierte Person eine maximal positive Bewertung der eigenen Kompetenz. Dabei sind an Lernzielen orientierte Personen eher ausdauernder, intrinsisch motivierter und engagierter in Lernaktivitäten als an Leistungszielen orientierte Personen (Wigfield & Cambria, 2010). In den 1990er-Jahren wurde zusätzlich zwischen der Annäherungskomponente (*approach component*) und der Vermeidungskomponente (*avoidance component*) innerhalb der zwei Zielorientierungen unterschieden (Elliot, 1999; Elliot & Harackiewicz, 1996). Das Annäherungsleistungsziel bezieht sich auf den Wunsch, die eigene Kompetenz zu demonstrieren und andere zu übertreffen, während das Vermeidungsleistungsziel den Zweck hat, nicht inkompetent im Vergleich zu andern Personen zu wirken. Annäherungsleistungsziele zeigen meist einen positiven Zusammenhang mit Leistung. Im Gegensatz dazu wirken Vermeidungsleistungsziele eher negativ auf Leistung (Wigfield & Cambria, 2010).

## 2.3 Erwartung-Wert-Theorien

Die Erwartung-Wert-Theorie integriert beide Ansätze der bisher genannten Motivationstheorien zu motivationalen Überzeugungen und den Gründen, warum sich Personen bei bestimmten Tätigkeiten engagieren. Bevor das moderne Erwartung-Wert-Modell beschrieben wird, soll der Vorläufer, das Risikowahl-Modell, vorgestellt werden.

### 2.3.1 Das Risikowahl-Modell

Das Risikowahl-Modell von Atkinson (1957, 1964) bildet eine wichtige Grundlage in der Forschung und gilt allgemeinhin als *das* Modell der Leistungsmotivation, da zum ersten Mal personale und situative Aspekte der Leistungsmotivation angemessen berücksichtigt wurden, wie sie auch im Grundmodell der Motivationspsychologie konzeptualisiert sind. Mit dem Modell wird die individuell wahrgenommene Aufgabenschwierigkeit auf Grundlage von Erwartung und Wert auf der Situationsseite (s. Abbildung 2.3) und den Motiven auf der Personenseite vorhergesagt. Erwartungen werden dabei als individuelle Beurteilungen der eigenen Fähigkeiten definiert und der Wert spiegelt die Wahrnehmung der Gründe wider, warum sich die Person anstrengen sollte. Auf der Personenseite befinden sich, wie oben schon erwähnt, die relativ zeitstabilen und überdauernden Motive, die in spezifischen Situationen ein bestimmtes Verhalten auslösen. In Leistungssituationen ist das Leistungsmotiv betroffen, das wiederum, je nach Richtung, in ein Annäherungsmotiv und ein Vermeidungsmotiv unterteilt werden kann: Ist eine Person in einer Leistungssituation eher erfolgszuversichtlich, überwiegt das Erfolgsmotiv ($M_e$); ist sie eher misserfolgsmeidend, herrscht das Misserfolgsmotiv vor ($M_m$). Da sowohl Hoffnung auf Erfolg als auch Furcht vor Misserfolg in einer Leistungssituation angeregt werden, ergibt sich aus der Differenz der beiden Motive die Nettohoffnung, die anzeigt, welches Motiv bei einer Person vorherrschend ist (Beckmann & Heckhausen, 2006; Brunstein & Heckhausen, 2006; Rheinberg, 2008).

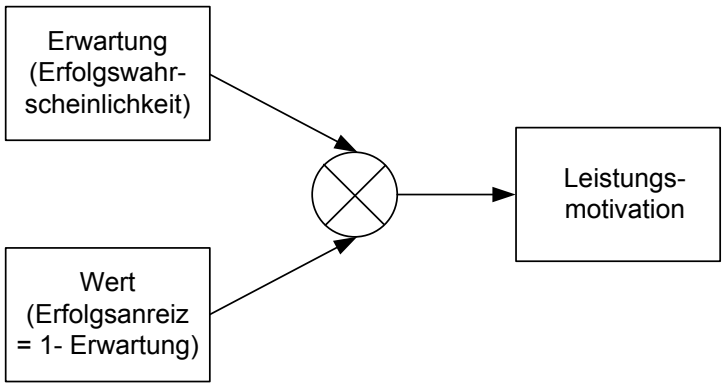



*Abbildung 2.3*. Risikowahl-Modell der Leistungsmotivation.

Ob nun ein Erfolgserleben empfunden wird, hängt von dem im Vorhinein gesetzten Anspruchsniveau ab, also dem Ziel, was ein Individuum plant, zu bewerkstelligen. Wenn das Vorhaben gelingt, wird die Person ein Erfolgserlebnis und damit Gefühle wie Freude

und Stolz erfahren; bei Misslingen wird die Person ein Misserfolgserlebnis empfinden einhergehend mit Gefühlen wie Ärger oder Scham. Dieses Anspruchsniveau wird wiederum von der Erfolgswahrscheinlichkeit (d. h. der Erwartung) und dem Erfolgsanreiz (d. h. dem Wert) beeinflusst. Korrespondierend zur Unterteilung des Leistungsmotivs wird auch hier zwischen Erfolgserwartung ($W_e$) beziehungsweise Misserfolgserwartung ($W_m$) sowie zwischen Erfolgsanreiz ($A_e$) beziehungsweise Misserfolgsanreiz ($A_m$) unterschieden. Dabei gilt

$$W_e = 1 - W_m \tag{1}$$

Das heißt, die Misserfolgserwartung ergibt sich als lineare Funktion der Erfolgserwartung, da sich Erfolg und Misserfolg gegenseitig ausschließen. Überdies sind Erwartung und Wert (d. h. der Anreiz) invers verknüpft, so dass sich die Formeln

$$A_e = 1 - W_e \tag{2}$$

und

$$A_m = 1 - W_m = - W_e \tag{3}$$

ergeben (Beckmann & Heckhausen, 2006). Ist eine Aufgabe sehr einfach, hat sie zwar eine hohe Erfolgswahrscheinlichkeit, da sie leicht zu lösen ist, aber der Anreiz geht deshalb gegen Null. Genau umgekehrt verhält es sich mit schwierigen Aufgaben, die zwar einen hohen Anreiz haben, aber die Erfolgswahrscheinlichkeit, sie zu lösen, geht gegen Null. Daher regen (subjektiv empfundene) mittelschwere Aufgaben das Leistungsmotiv am besten an, da hier sowohl Erfolg als auch Misserfolg möglich ist. Diese Feststellung untermauert auch Rheinberg, in dem er folgendes herausstellt: „Sie sind zwar anspruchsvoll, aber noch erreichbar und entsprechen am ehesten dem, was der Person mit vollem Einsatz gerade noch gelingt, ohne Anstrengung jedoch nicht“ (Rheinberg, 2008, S. 72). Mittelschwere Aufgaben stellen daher die meisten Informationen über die Individuen, also über ihre Anstrengung und Fähigkeit bereit (Schunk, Pintrich & Meece, 2008).

Diese Aussage trifft allerdings ausschließlich auf erfolgszuversichtliche Personen zu, bei denen das Leistungsmotiv vorherrschend ist, wie in Abbildung 2.4 dargestellt. Misserfolgsmeidende Personen hingegen, bei denen das Misserfolgsmotiv überwiegt, favorisieren entweder sehr leichte Aufgaben, weil diese quasi jeder schaffen kann, oder sehr schwierige Aufgaben, die de facto niemandem gelingen können. Denn bei diesen sehr leichten oder sehr schweren Aufgaben ist die Gefahr eines Misserfolges für sie am geringsten. Im Gegensatz zu erfolgszuversichtlichen Personen ist die Gefahr, zu scheitern, für misser-

folgsmeidende Personen bei mittelschweren Aufgaben am größten. Bei diesen Aufgaben sollen sie ihre Tüchtigkeit unter Beweis stellen, was für sie besonders bedrohlich ist, da sie einen Misserfolg antizipieren. In der Empirie konnte die Modellvorhersage für Erfolgszuversichtliche bestätigt werden (d. h. sie bevorzugten tatsächlich mittelschwere Aufgaben), jedoch nicht für Misserfolgsmotivierte, die keine bestimmten Aufgabenschwierigkeiten klar bevorzugten (Brunstein & Heckhausen, 2006; Rheinberg, 2008).

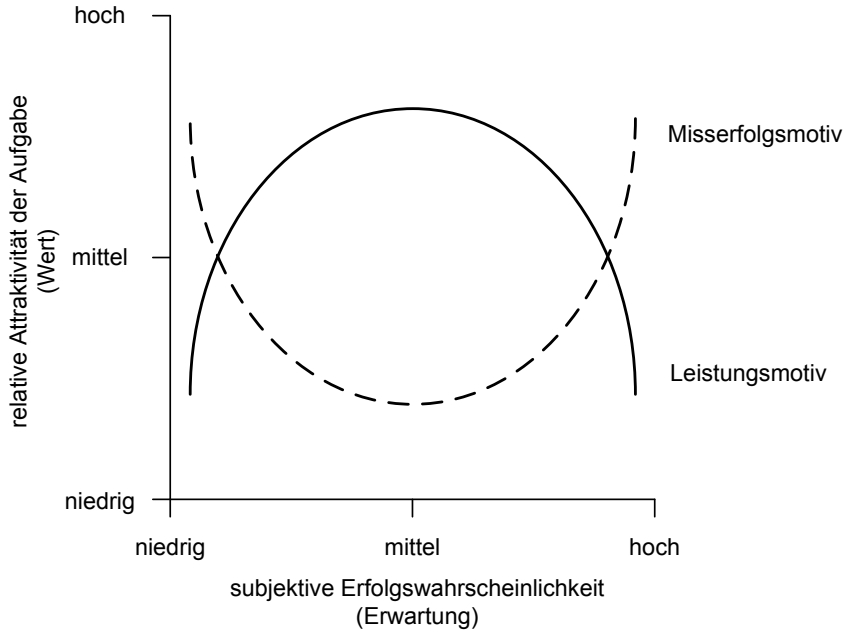



*Abbildung 2.4*. Die theoretische Kurve der Leistungsmotivation für erfolgszuversichtliche und misserfolgsmeidende Personen im Risikowahl-Modell (in Anlehnung an Atkinson, 1957, S. 365).

Mathematisch berechnet sich die Tendenz, Erfolg anzustreben, in dem das Erfolgsmotiv mit der Erfolgserwartung und dem Erfolgsanreiz multipliziert wird:

$$T_e = M_e \times W_e \times A_e \tag{4}$$

Analog ergibt sich die Tendenz, Misserfolg zu vermeiden, ebenfalls aus der Multiplikation des entsprechenden Motivs, der Erwartung und des Anreizes:

$$T_m = M_m \times W_m \times A_m \tag{5}$$

Welche Tendenz bei einer Person überwiegt ergibt sich aus der Summe der Erfolgs- und Misserfolgstendenz:

$$T_r = T_e + T_m = (M_e \times W_e \times A_e) + (M_m \times W_m \times A_m) \tag{6}$$

Durch das Einsetzen von Formel (3) in Formen (5) ergibt $T_m$ stets negative Werte, so dass es sich bei Formel (6) eigentlich um eine Differenzbildung handelt. Laut Risikowahl-Modell ist die resultierende Tendenz bei der Bearbeitung von mittelschweren Aufgaben am größten. Das bedeutet, überwiegt die Misserfolgstendenz in einer Leistungssituation ($T_e < T_m$), wird die Person die Situation meiden und eher sehr leichte oder sehr schwierige Aufgaben bearbeiten. Überwiegt hingegen die Erfolgstendenz ($T_e > T_m$), wird die Person sich der Situation annähern und vor allem Aufgaben mittlerer Schwierigkeit bearbeiten (Asseburg, 2011; Beckmann & Heckhausen, 2006).

Zusammenfassend entsteht im Risikowahl-Modell leistungsmotiviertes Verhalten durch eine multiplikative Verknüpfung von Erwartung und Wert. Das bedeutet, wenn sich eine Person zwar fähig fühlt, eine Aufgabe zu meistern, diese aber nicht wertschätzt, ist es unwahrscheinlich, dass sie sich anstrengt. Dies gilt ebenso für die andere Richtung: Wenn eine Person eine Aufgabe wertschätzt, sich aber unfähig fühlt, diese erfolgreich zu bearbeiten, ist es kaum denkbar, dass eine hohe Anstrengung investiert wird (Schunk et al., 2008).

Ursprünglich wurde das Risikowahl-Modell nur auf die Auswahl von Aufgaben angewendet; der Geltungsbereich wurde jedoch auch auf andere leistungsthematische Konstrukte wie Anstrengung, Ausdauer und Leistung erweitert. Bezüglich der Aufgabenauswahl und Ausdauer konnte das Modell bestätigt werden, für den Zusammenhang mit Leistung jedoch nicht. Zwar hängt die Quantität der bearbeiteten Aufgaben auch von der Motivation ab, aber ob das auch für die Qualität der Leistung gilt, bleibt unklar (Beckmann & Heckhausen, 2006).

Auch wenn das Risikowahl-Modell als das zentrale Modell der Leistungsmotivation gilt, birgt es konzeptuelle Nachteile. Zum einen wurde die multiplikative Verknüpfung von Erfolgswahrscheinlichkeit und Erfolgsanreiz in Studien widerlegt, die eine positive Korrelation der zwei Komponenten konstatierten (Schunk et al., 2008; Wigfield & Eccles, 1992). Auch die Konzeptualisierung des Wertes einer Aufgaben als lineare Funktion der Erfolgserwartung kam im Risikowahl-Modell zu kurz und wurde in nachfolgenden Modellen von Eccles und Wigfield (s. Abschnitt 2.3.2) behoben. Ebenfalls ist problematisch, dass die Annahmen des Modells nur für erfolgszuversichtliche Personen validiert werden konnten (Asseburg, 2011; Schunk et al., 2008). Die aufgezeigten Nachteile haben zu mehreren Veränderungen des Risikowahl-Modells und damit zur Entwicklung des Erwartung-Wert-Modells der Leistungsmotivation geführt.

### 2.3.2 Das Erwartung-Wert-Modell der Leistungsmotivation

Leistungsmotivation als aktueller Zustand bildet den Ausgangspunkt der Erwartung-Wert-Theorien. Basierend auf dem Risikowahl-Modell von Atkinson (1957, 1964) erklärt das Erwartung-Wert-Modell (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000) leistungsmotiviertes Verhalten durch das Zusammenwirken von Erwartungen und Werten, wobei Motive nur noch eine untergeordnete Rolle spielen. Wie in Abbildung 2.5 verdeutlicht, wird angenommen, dass die Erwartungen und Werte direkt einen Einfluss auf die Entscheidung für eine Tätigkeit besitzen sowie auf die Leistung, die investierte Anstrengung und Ausdauer. Dabei fließt in diesem, um die soziokulturelle Umwelt der Individuen erweiterten Modell die Vorhersage der Leistung ausdrücklich mit ein. Die Wertkomponente wird als eigenständige Komponente konzeptualisiert und nicht mehr nur als lineare Funktion der Erfolgserwartung definiert (Schunk et al., 2008).

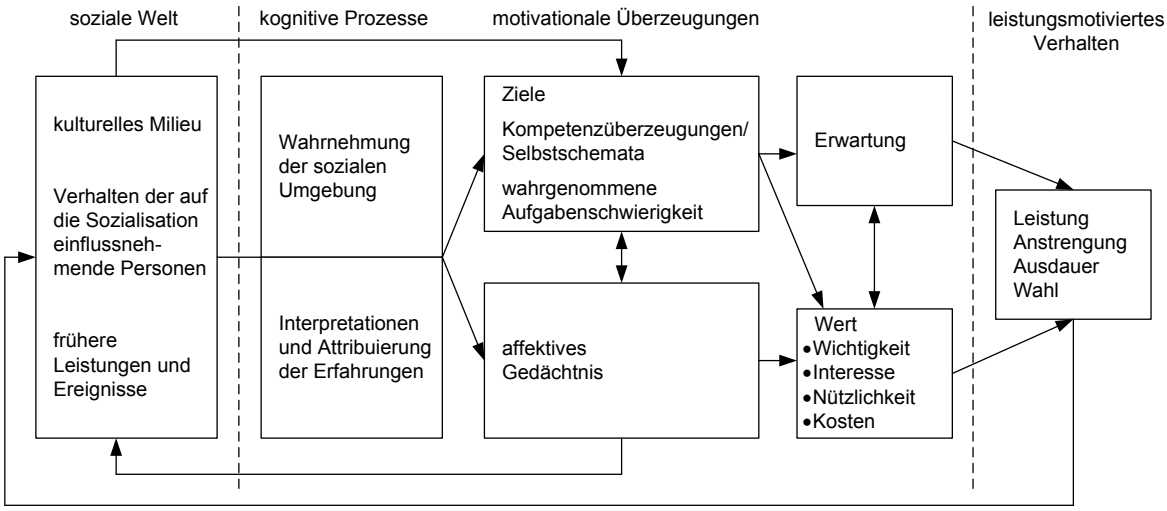



*Abbildung 2.5*. Erwartung-Wert-Modell der Leistungsmotivation (übersetzt von Schunk et al., 2008, S. 51; die Korrelation zwischen Erwartung und Wert fehlt in der Originalabbildung, sie ist jedoch im Erwartung-Wert-Modell vorgesehen; Eccles & Wigfield, 2002).

Die Erwartungskomponente beschreibt die Vorstellungen der Schülerinnen und Schüler darüber, wie gut sie in der zu bearbeitenden Aufgabe abschneiden werden, und stellt die individuelle Wahrnehmung der eigenen aktuellen Kompetenz zu einer gegebenen Aktivität dar. Die Personen stellen sich die Frage: „Bin ich fähig, diese Aufgabe zu bewältigen?“. Dabei ist die Erwartungskomponente konzeptuell eng mit domänenspezifischen Kompetenzüberzeugungen (z. B. Selbstkonzept) verknüpft, wobei diese eher auf die gegenwärtige Fähigkeit in einer Domäne ausgerichtet sind und die Erfolgserwartungen auf

bevorstehende Aufgaben und Tätigkeiten. Die Differenzierung von Erfolgserwartungen und Kompetenzüberzeugungen konnte empirisch allerdings nicht belegt werden, da Schülerinnen und Schüler beispielsweise nicht zwischen ihren mathematischen Fähigkeiten und ihrer Erwartung bezüglich des Abschneidens in einem Mathematiktest unterscheiden können (Schunk et al., 2008; Wigfield & Eccles, 2002). Erfolgserwartungen werden in ähnlicher Weise wie die Wirksamkeitserwartungen des Selbstwirksamkeitskonstrukts gemessen und korrespondieren daher konzeptuell mit domänenspezifisch erfassten Wirksamkeitserwartungen (Eccles & Wigfield, 2002).

Die Wertkomponente hingegen erklärt leistungsbezogene Entscheidungen anhand der Überzeugungen der Schülerinnen und Schüler, warum sich die Aufgabenbearbeitung für sie lohnen sollte. Die korrespondierende Frage lautet: „Warum sollte ich diese Aufgabe bearbeiten?“. Dabei wird zwischen den vier Wertaspekten Wichtigkeit, Interesse, Nützlichkeit sowie Kosten differenziert, wobei jede Komponente das Leistungsverhalten beeinflussen kann. Diese Aspekte des subjektiven Aufgabenwertes beziehen sich auf individuelle Wahrnehmungen des Wertes und des Interesses an der Tätigkeit. Eine Aufgabe kann mehrere Quellen des Wertes haben, wobei gilt, je mehr Quellen vorhanden sind, umso höher ist der wahrgenommene Wert (Eccles, 2005). Die Wertaspekte wirken gleichzeitig, wodurch der Wert einer Aufgabenbearbeitung festgelegt wird (Eccles & Wigfield, 2002; Schunk et al., 2008; Wigfield & Eccles, 2000). Im Folgenden werden die verschiedenen Wertaspekte definiert und Parallelen zu den in Abschnitt 2.2.2 vorgestellten Konstrukten gezogen. Beim Vergleich der unterschiedlichen Konstrukte der Erwartung-Wert-Theorie mit anderen motivationalen Konstrukten ist allerdings zu beachten, dass sie jeweils aus unterschiedlichen theoretischen Traditionen stammen und ihnen daher auch verschiedene Bedeutungen immanent sind.

Der Wichtigkeitsaspekt (*attainment value*) drückt die individuelle wahrgenommene Bedeutung eines guten Abschneidens bei der aktuellen Tätigkeit aus. Die Wichtigkeit kennzeichnet das Ausmaß, zu welchem die Aufgabe den Individuen ermöglicht, zentrale Aspekte ihrer Selbstschemata zu bestätigen. Eine Person wertet eine Aufgabe demnach dann als wichtig, wenn diese als zentral für das eigene Selbst angesehen wird (Eccles & Wigfield, 2002). Beispielsweise engagiert sich eine Person bei einer Aufgabe, da sie durch diese ihre Gender-Rolle bestätigten kann. Das Selbstschema ist ein Teil unserer sozialen und persönlichen Identität, die einen starken Einfluss darauf nehmen, wie verschiedene Möglichkeiten gewertet werden (Eccles, 2005).

Der Interessenaspekt (*intrinsic value*) bezieht sich auf die Freude während der Tätigkeit und beinhaltet situative sowie andauernde Aspekte, so dass es dem Interessenkonstrukt (s. Abschnitt 2.2.2) ähnelt, das in individuelles und situatives Interesse unterteilt wird. Wenn Personen Freude an einer Tätigkeit haben, sind sie meist engagiert und ausdauernd bei der Sache. Dieses Verhalten ähnelt den Beschreibungen zum Konstrukt des individuellen Interesses. Da die Wertkomponente generell den subjektiven Wert von Aufgaben beschreibt, ist es allerdings auch vorstellbar, dass der Interessenaspekt von Situation zu Situation variiert. Dies korrespondiert wiederum mit dem Konstrukt des situativen Interesses. Zusammenfassend zeigt der Interessensaspekt eine große Situationsspezifität und ist demnach mehr mit der Ausführung der Tätigkeit verbunden sowie mit Freude bei der Durchführung als mit dem Ergebnis der Tätigkeit. Wenn das Interesse hoch ist, dann strengen sich die Personen eher an und sind ausdauernder bei der Aufgabenbearbeitung (Schunk et al., 2008; Wigfield & Cambria, 2010; Wigfield & Eccles, 2000).

Nützlichkeit (*utility value*) beschreibt die Passung der aktuellen Aufgabe mit den zukünftigen Plänen des Individuums. Daher bezieht sich dieser Wertaspekt konzeptuell eher auf das Ergebnis einer Aufgabe. Dabei kann eine Tätigkeit als nützlich wahrgenommen werden, wenn sie mit eigenen wichtigen Zielen korrespondiert. In diesem Sinne weist der Nützlichkeitsaspekt Ähnlichkeiten mit den erwähnten Zielorientierungen auf (Eccles & Wigfield, 2002; Schunk et al., 2008; Wigfield & Cambria, 2010).

Die Kosten (*cost*) entsprechen Opportunitätskosten der Aufgabenbearbeitung, also inwiefern die Entscheidung für eine Aufgabe die Ausführung andere Tätigkeiten einschränkt. Darüber hinaus kann die antizipierte oder investierte Anstrengung hier verortet werden, die für eine erfolgreiche Aufgabenbearbeitung notwendig ist. Die Kosten sind vor allem für Wahlentscheidungen von Bedeutung, da jede Entscheidung für eine Tätigkeit gleichzeitig eine Entscheidung gegen andere Tätigkeiten bedeutet. Allerdings wurde diesem Aspekt der Wertkomponente bisher am wenigsten Aufmerksamkeit in der Forschung geschenkt (Wigfield & Cambria, 2010; Wigfield & Eccles, 2000).

Die einzelnen Wertaspekte des Erwartung-Wert-Modells sollten nicht isoliert betrachtet werden, sondern immer im Zusammenspiel mit den anderen Wertaspekten, wenn auf Leistungsverhalten rückgeschlossen wird (Wigfield & Cambria, 2010). Denn alle vier Wertaspekte bilden zusammen den subjektiven Wert einer Aufgabe (Eccles, 2005). Neben den beschriebenen Beziehungen zwischen den Konstrukten wäre es ebenfalls möglich, Querweise zu einer Vielzahl anderer Konstrukte der Motivationsforschung zu ziehen wie

zum Beispiel zwischen Wichtigkeit und integrierter Regulation der Selbstbestimmungstheorie oder zwischen Interesse und intrinsischer Motivation (Ryan & Deci, 2000), worauf an dieser Stelle jedoch verzichtet wird.

Die Überprüfung der Gültigkeit einzelner Komponenten der Erwartung-Wert-Theorie kam zu dem Ergebnis, dass Selbstkonzept, Selbstwahrnehmung der Fähigkeit und Erfolgserwartung keine separaten Faktoren bilden, sondern auf einem gemeinsamen Faktor laden. Obwohl auf theoretischer Ebene das Selbstkonzept der Fähigkeit allgemeiner und stabiler ist als die situationsspezifischen Erfolgserwartungen, scheinen Jugendliche diese Konstrukte bei der Beantwortung von Fragebögen nicht zu unterscheiden. Hingegen bildeten die drei Wertkomponenten Wichtigkeit, Interesse und Nützlichkeit empirisch distinkte Faktoren ab. Erwartung und Wert sind positiv korreliert, so dass bezüglich ihrer Fähigkeits- und Erfolgserwartung selbstbewusste Jugendliche, die ihre Fähigkeiten als hoch einschätzen, dazu tendieren, die Aufgabe auch positiv zu werten (Eccles & Wigfield, 1995).

Forschung zum Erwartung-Wert-Modell konnte zeigen, dass die Erwartungskomponente (inklusive Erfolgserwartungen, Selbstkonzept und anderen Kompetenzüberzeugungen) vor allem die Leistung wie die Note oder Testleistung in standardisierten Leistungstests vorhersagt, aber auch die Anstrengungsbereitschaft. Dabei wiesen die Erfolgserwartungen eine höhere Vorhersagekraft für spätere Schulnoten in Mathematik und Englisch auf als bisherige Schulnoten in diesen Fächern, auch nach Kontrolle von vorheriger Leistung. Die Wertkomponente ist ebenfalls positiv mit Leistung korreliert. Werden aber Erwartung und Wert gleichzeitig als Prädiktoren für Leistung angewendet, zeigt nur die Erwartungskomponente eine signifikante Beziehung mit Leistung (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000). Die einzelnen Wertaspekte sind bessere Prädiktoren für Wahlentscheidungen (z. B. Leistungskurswahl) oder Ausdauer bei begonnener Aufgabenbearbeitung (Eccles & Wigfield, 2002; Schunk et al., 2008; Wigfield & Eccles, 2000, 2002).

Wie in Abbildung 2.5 ersichtlich, werden neben diesen zwei Hauptfaktoren, Erwartung und Wert, zahlreiche andere Faktoren berücksichtigt, die für die Entwicklung von Erwartung und Wert bedeutsam sind. Dazu gehören aufgabenspezifische Vorstellungen wie Kompetenzüberzeugungen (z. B. domänenspezifisches Selbstkonzept), die wahrgenommene Schwierigkeit einer Aufgabe, individuelle Ziele (kurz- oder langfristig) und Selbstschemata. Affektive Erinnerungen beziehen sich auf vorangegangene Erfahrungen

mit solchen Aufgaben. Diese sozial-kognitiven Komponenten werden wiederum durch die subjektive Wahrnehmung der individuellen Umgebung und die Interpretation und Ursachenzuschreibung vorangegangener Erfahrungen beeinflusst. Ebenfalls wirken die soziokulturelle Umwelt und eine Vielzahl von Sozialisationseinflüssen. Die motivationalen Überzeugungen sind demnach in einem großen sozialen und kulturellen Kontext verortet und werden durch sozial-kognitive Prozesse konstruiert (Eccles & Wigfield, 2002; Schunk et al., 2008; Wigfield & Eccles, 2000). Insgesamt berücksichtigt das Modell sowohl Personen- als auch Situationsmerkmale, wie es im Grundmodell der klassischen Motivationspsychologie konzeptualisiert ist.

Basierend auf dem Erwartung-Wert-Modell definiert der nächste Abschnitt das Konstrukt der Testteilnahmemotivation und verortet es im Kontext der Leistungsmotivation. Anschließend wird der Forschungsstand zu Testteilnahmemotivation umrissen und ein aktuelles Modell für die Anstrengungsbereitschaft erläutert, bevor ein Erwartung-Wert-Anstrengung-Modell der Testteilnahmemotivation abgeleitet wird.

## 2.4 Testteilnahmemotivation

### 2.4.1 Definition und Verortung

Baumert und Demmrich (2001) definieren Testteilnahmemotivation als „the willingness to engage in working on test items and to invest effort and persistence in this undertaking" (S. 441). Eine Person gilt demnach in einer Testsituation als motiviert, wenn ihr Ziel eine erfolgreiche Aufgabenbearbeitung ist und sie dieses Ziel durch permanente Anstrengung versucht, zu erreichen. Damit lässt sich Testteilnahmemotivation der aktuellen Motivation im Allgemeinen und der situationsspezifischen Leistungsmotivation im Besonderen zuordnen, deren Entstehung im Grundmodell der Motivationspsychologie (s. Abschnitt 2.1.3) und im Erwartung-Wert-Modell (s. Abschnitt 2.3.2) thematisiert wird. Die Situationsspezifik bezieht sich im Low-Stakes-Kontext auf die Besonderheit, dass der Test keine Konsequenzen für die Schülerinnen und Schüler nach sich zieht, unabhängig davon, wie erfolgreich beziehungsweise erfolglos sie die Aufgaben bearbeiten (Eklöf, 2010a). Damit kommt der Anstrengungsbereitschaft der Schülerinnen und Schüler eine besondere Rolle zu.

Wird das Grundmodell der „klassischen" Motivationspsychologie auf Testteilnahmemotivation angewendet, gelten folgende Annahmen: Leistungsmotiviertes Verhalten in einem Test wird durch die aktuelle Motivation zur Testbearbeitung, also der Testteilnah-

memotivation unmittelbar beeinflusst; Personen- und Situationsmerkmale wirken indirekt auf das Verhalten. Im Kontext von Schulleistungsstudien ohne Konsequenzen für die Testteilnehmenden ist die Situationskomponente besonders ausschlaggebend. In solch einer Situation scheint es schwieriger, dass das Leistungsmotiv der Schülerinnen und Schüler angeregt wird, als zum Beispiel in Schultests, deren Ergebnisse benotet werden. Auf Basis des Grundmodells sollten bei der Untersuchung von Testteilnahmemotivation sowohl überdauernde, stabile Personeneigenschaften (sogenannte *traits*) als auch situationsspezifische, instabile Merkmale (*states*) berücksichtigt werden (Asseburg, 2011). Testteilnahmemotivation als ein besonderer Zustand (d. h. *state*) der *trait*-ähnlichen Leistungsmotivation kann in bestimmten Situationen von der Leistungsmotivation divergieren, so dass eine differenzierte Betrachtung von *trait* und *state* relevant ist. Beispielsweise ist eine Schülerin besonders gut in Mathematik und hat eine Vorliebe für dieses Fach; sie bearbeitet dennoch einen Low-Stakes-Assessment in Mathematik unmotiviert, da sie keine positiven oder negativen Konsequenzen zu erwarten hat.

Als theoretische Basis wird oft das Erwartung-Wert-Modell (Wigfield & Eccles, 2000) angewendet, das für die Konzeptualisierung der Testteilnahmemotivation durch die Erklärung des Zusammenhangs zwischen Motivation und Testleistung besonders geeignet ist. Wie schon in der Problemstellung erwähnt, ist es ungewiss, ob sich die Schülerinnen und Schüler in Low-Stakes-Assessments anstrengen und ihre beste Leistung zeigen. In Anwendung des Erwartung-Wert-Modells kann gefragt werden, *warum* die Teilnehmenden den Test wertschätzen (ihn also als wichtig, nützlich oder interessant wahrnehmen) und *warum* sie sich maximal anstrengen sollten, wenn das Testergebnis keine persönlichen Auswirkungen hat. Der inhärent niedrige Wert eines solchen Tests könnte erklären, warum Schülerinnen und Schüler geringe Anstrengungsbereitschaft berichten und wiederum eine geringere Testleistung aufweisen könnten als in High-Stakes-Tests. Wird Testteilnahmemotivation im Kontext der Erwartung-Wert-Theorie konzeptualisiert, wie es in dieser Arbeit geschieht, dann ist zu beachten, dass es sich um ein mehrdimensionales Konstrukt handelt (Sundre, 2007), das, neben der in der Definition explizit genannten Anstrengungsbereitschaft, ebenfalls die Erfolgserwartungen und den wahrgenommenen Wert eines Tests umfasst.

Derzeit gibt es kein fundiertes Modell zur aktuellen Testteilnahmemotivation, das alle drei Komponenten (Erwartung, Wert und Anstrengung) berücksichtigt. Auch fehlen Motivationsmodelle, die nicht nur motiviertes Verhalten wie Ausdauer und Anstrengung,

sondern auch Leistung vorhersagen (Rheinberg, 2008). Bevor ein Erwartung-Wert-Anstrengung-Modell der Testteilnahmemotivation entworfen wird (Abschnitt 2.4.4), wird im Folgenden ein Modell zur Anstrengungsbereitschaft (Abschnitt 2.4.2) vorgestellt und der empirische Forschungsstand zur Testteilnahmemotivation (Abschnitt 2.4.3) aufgezeigt.

### 2.4.2 Das demands-capacity model of test-taking effort

Die Definition von Testteilnahmemotivation demonstriert die Wichtigkeit der Anstrengungsbereitschaft in Low-Stakes-Assessments als eine zentrale Komponente der Testteilnahmemotivation. Wise und Smith (2011) entwickelten auf theoretischer Ebene ein Modell für die Anstrengungsbereitschaft von Testteilnehmenden, das *demands-capacity model of test-taking effort*. Dies wird an dieser Stelle erläutert, da es die dynamischen Prozesse der Testteilnahmemotivation während der Testsitzung explizit in den Fokus nimmt und zwischen anfänglicher Anstrengung und Anstrengung im Verlauf eines Tests differenziert.

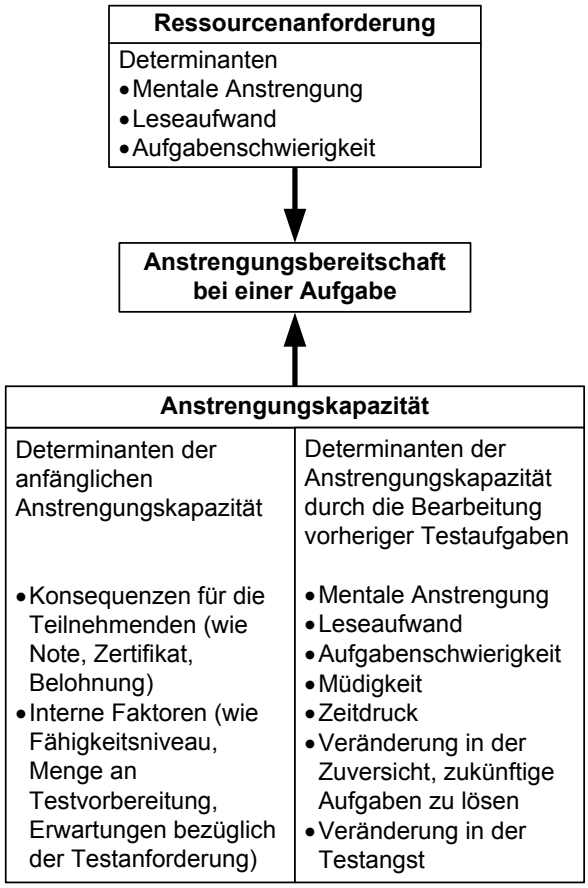



*Abbildung 2.6*. Das *demands-capacity model of test-taking effort* (in Anlehnung an Wise & Smith, 2011, S. 149; eigene Übersetzung).

Wie in Abbildung 2.6 verdeutlicht, wird in diesem Modell eine Vielzahl von Faktoren mit Anstrengungsbereitschaft für eine Aufgabe in Relation gesetzt. Dabei bilden die Ressourcenanforderung (*resource demands*) und die Anstrengungskapazität (*effort capacity*) die beiden Hauptdeterminanten der Anstrengungsbereitschaft. Die Ressourcenanforderung ist als Aufgabenmerkmal definiert, das die aufzuwendende Anstrengung repräsentiert, die für die Lösung der Aufgabe notwendig ist, wie zum Beispiel die Aufgabenschwierigkeit oder der Leseaufwand. Diese Anforderung kann zwischen den Aufgaben variieren. Die Anstrengungskapazität ist ein Merkmal auf der Personenseite und beschreibt das Maß an Anstrengung, das die Teilnehmenden bereit sind, in die Lösung der Aufgaben zu investieren. Diese Kapazität kann sowohl zwischen einzelnen Personen als auch zwischen Aufgaben variieren. Das Modell korrespondiert durch die Berücksichtigung von Situations- beziehungsweise Aufgabenmerkmalen und Personenmerkmalen mit dem Grundmodell der klassischen Motivationspsychologie (Rheinberg, 2008).

Die Anstrengungskapazität der Testteilnehmenden wird unterteilt in a) anfängliche Anstrengungskapazität, die zum Beispiel von der Testkonsequenz abhängt, und in b) Anstrengungskapazität im Verlauf des Tests, die durch die Bearbeitung vorheriger Aufgaben beeinflusst wird. Auswirkungen auf Anstrengungskapazität im Verlauf des Tests haben unter anderem Ermüdungserscheinungen durch sehr herausfordernde, vorangegangene Aufgaben, Testangst oder Veränderung im Selbstvertrauen, zukünftige Aufgaben zu lösen. Damit werden auch individuelle Veränderungen der Anstrengungsbereitschaft während des Verlaufs eines Tests in diesem Modell betrachtet und somit die motivationalen Prozesse, die der Bearbeitung eines Leistungstests unterliegen. Darüber hinaus werden Erfolgserwartungen (wahrgenommene Aufgabenschwierigkeit, Veränderung des Selbstvertrauen) und Wertaspekte (Testangst, Herausforderungsgrad der Aufgaben) neben anderen Faktoren der Anstrengungsbereitschaft berücksichtigt (Wise & Smith, 2011). Mit diesem speziell für Anstrengungsbereitschaft im Kontext von Low-Stakes-Assessments entwickelten Modell wird deutlich, von welchen Determinanten die intendierte und auch investierte Anstrengungsbereitschaft der Teilnehmenden bei der Bearbeitung eines Tests abhängt. Außerdem wird die Veränderbarkeit von Testteilnahmemotivation während einer Testsitzung betont, was im Folgenden noch relevant wird. Bevor das Kapitel zur Fragestellung anschließt, wird der Forschungsstand zusammengefasst, um die Lücken beziehungsweise Unklarheiten in diesem zu verdeutlichen, und ein Erwartung-Wert-Anstrengung-Modell skizziert.

### 2.4.3 Empirischer Forschungsstand zu Testteilnahmemotivation

Dieser Abschnitt beschäftigt sich mit den empirischen Ergebnissen, die Studien im Bereich Testteilnahmemotivation ergaben. Zunächst wird auf den Unterschied in der Motivation zwischen Low- und High-Stakes-Test eingegangen. Anschließend werden Ergebnisse aus Studien dargestellt, die auf der Erwartung-Wert-Theorie basieren, jedoch eine theoretische Komponente des Modells (entweder Erwartung oder Anstrengungsbereitschaft) außer Acht lassen. Es folgt der Forschungsstand zur Unterscheidung zwischen domänen- und situationsspezifischer Leistungsmotivation, um im Anschluss auf Studien einzugehen, die die Veränderung der Testteilnahmemotivation während einer Testsitzung untersuchten.

*Motivationale Unterschiede zwischen Low- und High-Stakes-Tests*

Ein bedeutender Strang der Testteilnahmemotivationsforschung befasst sich mit Unterschieden in der Motivation und Leistung von Teilnehmenden in Low- und High-Stakes-Tests. Dabei ziehen High-Stakes-Tests, im Gegensatz zu Low-Stakes-Tests, Konsequenzen nach sich, wie zum Beispiel die Benotung des Testergebnisses. Wolf und Smith (1995) führten eine experimentelle Studie durch, in der einer Gruppe der Testteilnehmenden mitgeteilt wurde, dass ihr Ergebnis benotet wird (*high-stakes*); die Ergebnisse der anderen Gruppe wurden nicht benotet (*low-stakes*). Die Testteilnahmemotivation wurde mit Fragen zur investierten Anstrengungsbereitschaft und zur Wichtigkeit des Tests operationalisiert. Es wurden große motivationale Unterschiede zwischen den Gruppen mit einer Effektstärke von knapp eineinhalb Standardabweichungen zugunsten der High-Stakes-Gruppe gefunden. Der Unterschied in der Gesamtleistung fiel zum Vorteil der High-Stakes-Gruppe mit einem Leistungsvorsprung von einer Viertel Standardabweichung aus. In einer anderen Studie von Sundre und Kitsantas (2004) zeigte eine Gruppe, die den Test ohne eine Benotung bearbeitete, ebenfalls eine niedrigere Motivation als die Gruppe mit einer Benotung des Testergebnisses. Auch hier wurde Motivation anhand von Fragen zur investierten Anstrengungsbereitschaft und Wichtigkeit des Tests gemessen. Die Testteilnahmemotivation war zudem nur in der Gruppe, die den Low-Stakes-Test bekamen, ein signifikanter Prädiktor der Testleistung. Unterschiede in der Testleistung und Testreliabilität zwischen High- und Low-Stakes-Tests untersuchten auch Cole und Osterlind (2008). Das Testergebnis der High-Stakes-Gruppe war die Zugangsvoraussetzung für ein Studienfach. Unter Berücksichtigung der Vorleistung und des Geschlechts der Teilnehmenden fanden die Autoren den größten signifikanten Unterschied in der Leistung zwischen der Low-Stakes- und High-Stakes-Bedingung für den Mathematiktest, auch wenn die Effekt-

stärke eher klein war. Insgesamt zeigte eine Synthese von zwölf Studien, dass die motivierten Testteilnehmenden (High-Stakes-Gruppe) im Durchschnitt um über eine halbe Standardabweichung bessere Ergebnisse erzielten als die unmotivierten Testteilnehmenden (Low-Stakes-Gruppe; Wise & DeMars, 2005). Erwin und Wise (2002) bringen es mit der folgenden Aussage auf den Punkt: „(…) the challenge to motivate our students to give their best effort when there are few or no personal consequences is probably the most vexing assessment problem we face" (S. 71).

Als bedeutende nationale Untersuchung ist die Studie von Baumert und Demmrich (2001) zu nennen, in der eine Kurzform des PISA-Mathematiktests eingesetzt wurde, um zu ermitteln, ob die Höhe der Testteilnahmemotivation und Leistung durch eine Erhöhung der Stakes verbessert werden kann. Es wurden vier Bedingungen unterschieden: 1) informatorisches Feedback, 2) Benotung, 3) finanzielle Belohnung und 4) Kontrollgruppe, die die übliche PISA-Instruktion erhielt, in der die gesellschaftliche Wichtigkeit der Testteilnahme an internationalen Vergleichsstudien betont wurde. Die Testteilnahmemotivation wurde durch Fragen a) zum persönlichen Wert, gut abzuschneiden, b) zur wahrgenommenen Nützlichkeit der Testteilnahme, c) zur intendierten und investierten Anstrengung sowie d) zu aufgabenirrelevanten Kognitionen operationalisiert. In allen vier Gruppen war die intendierte und investierte Anstrengungsbereitschaft hoch bis sehr hoch und es wurden keine Effekte auf die Testleistung gefunden. Ebenfalls ergaben sich keine Unterschiede im persönlichen Wert einer erfolgreichen Testbearbeitung und in der wahrgenommenen Nützlichkeit des Tests zwischen die Gruppen. Zusätzlich wurden Differenzen zwischen Schülerinnen und Schülern der Hauptschule und Gymnasiastinnen und Gymnasiasten untersucht. Die *intendierte* Anstrengungsbereitschaft unterschied sich nicht zwischen den Gruppen; allerdings berichteten Teilnehmende der Hauptschule, dass sie tatsächlich weniger Anstrengung *investierten* als ihre Schulkameradinnen und Schulkameraden auf dem Gymnasium. Darüber hinaus berichteten die Teilnehmenden, die ein Gymnasium besuchten, einen positiveren emotionalen Zustand sowie weniger Sorgen und Ablenkung als Jugendliche an einer Hauptschule (Baumert & Demmrich, 2001).

*Testteilnahmemotivation zwischen Erwartung, Wert und Anstrengungsbereitschaft*

In der Literatur existiert bisher kein eigenes theoretisches Modell, das Testteilnahmemotivation konzeptualisiert. Die meisten empirischen Studien stützen sich allerdings auf die Erwartung-Wert-Theorie (Wigfield & Eccles, 2000). Durch die explizite Berücksichtigung von Leistung und Anstrengung im Erwartung-Wert-Modell ist dieses situativ ausgerichtet

und eignet sich auch für die Konzeptualisierung von Testteilnahmemotivation (Asseburg, 2011; Schunk et al., 2008). Obwohl empirisch belegt werden konnte, dass die Erwartungskomponente ein stärkerer Prädiktor für Leistung im Vergleich zur Wertkomponente ist, berücksichtigten die meisten Studien ausschließlich die Wertkomponente und Anstrengungsbereitschaft als Operationalisierung der Testteilnahmemotivation (Cole, Bergin & Whittaker, 2008; Eklöf & Nyroos, 2013; Eklöf, Pavešič & Grønmo, 2014; Swerdzewski, Harmes & Finney, 2011; Wolf & Smith, 1995). Einige wenige Untersuchungen, die sowohl die Erwartungs- als auch die Wertkomponente berücksichtigten, erhoben keine Anstrengungsbereitschaft, die eine Schlüsselkomponente in der Definition von Testteilnahmemotivation darstellt (Asseburg, 2011; Freund & Holling, 2011; Freund, Kuhn & Holling, 2011). Bisher hat also keine Studie, die theoretisch auf dem Erwartung-Wert-Modell basiert, sowohl *Erwartung* und *Wert* als auch *Anstrengungsbereitschaft* in ihrer Analyse berücksichtigt.

In der Studie von Asseburg (2011), die die Erwartung und den Wert einer erfolgreichen Testbearbeitung von Schülerinnen und Schülern der neunten Klasse untersuchte, zeigte die Wertkomponente *keinen* signifikanten Einfluss auf die Mathematikleistung. Daher empfiehlt sie, von einem „Erwartung-Modell der Motivation zur Testbearbeitung“ zu sprechen (Asseburg 2011, S. 111). Die Wertkomponente spielt demnach eher eine untergeordnete Rolle für die Erklärung von Testleistung. Dies korrespondiert mit der Aussage von Wolf und Smith (1995), dass die Wichtigkeit einer guten Leistung, die einen Aspekt der Wertkomponente darstellt, von der Testkonsequenz abhängt. Da das Testergebnis bei Low-Stakes-Tests keinerlei Folgen für die Testteilnehmenden hat, scheint die Wertkomponente solcher Tests per se niedrig zu sein.

Asseburgs Empfehlung, von einem Erwartung-Modell auszugehen, wird von Cole und Kollegen (2008) kontrastiert, die die Bedeutung der Wertkomponente für Low-Stakes-Tests explizit betonen und im Gegenzug die Erwartungskomponente für vernachlässigbar halten. Ihrer Einschätzung nach kann im Low-Stakes-Kontext „Erfolg“ der Erwartungskomponente nicht definiert werden, da die Testteilnehmenden nicht ihre erreichte Punktzahl erfahren oder alternative Rückmeldungen bekommen. Daher betrachteten sie in ihrer Studie mit Studierenden ausschließlich die Wertkomponente und erhoben Interesse, Nützlichkeit, Wichtigkeit sowie Anstrengung. Die Ergebnisse der Pfadanalysen für die Mathematikleistung zeigten, dass Wichtigkeit und Nützlichkeit starke Prädiktoren der Anstrengungsbereitschaft waren und Anstrengungsbereitschaft wiederum Testleistung

vorhersagte unter Kontrolle von Geschlecht und Vorleistung. Dabei wurden die Effekte von Wichtigkeit und Nützlichkeit auf Leistung vollständig durch Anstrengungsbereitschaft mediiert; der Effekt von Interesse auf Leistung wurde partiell mediiert (Cole et al., 2008). Auch Zilberberg, Finney, Marsh und Anderson (2014) konnten die mediierende Rolle von Anstrengungsbereitschaft replizieren: In ihrer Untersuchung sagte die wahrgenommene Wichtigkeit des Tests signifikant Anstrengungsbereitschaft vorher; Anstrengungsbereitschaft wiederum zeigte einen signifikanten positiven Zusammenhang mit Testleistung auf unter Kontrolle von Geschlecht und Problemlösefähigkeit. Diesen Studien zeigen, dass Anstrengungsbereitschaft den Effekt der Wertkomponente auf Leistung vermittelt. Ob dies auf die Erwartungskomponente übertragen werden kann, ist bisher unerforscht.

Die Erwartungskomponente wurde in deutschen Studien häufiger erfasst, die die Motivation von Testteilnehmenden untersuchten. So wurde im nationalen Kontext auf Basis des Fragebogens zur Erfassung der aktuellen Motivation (Rheinberg, Vollmeyer & Burns, 2001) die Testteilnahmemotivation in Low-Stakes-Assessments *vor* dem Leistungstest erfasst. Freund et al. (2011) erhoben die Erfolgserwartung und diverse Wertaspekte bevor die Testteilnehmenden einen Test zum abstrakten Denken bearbeiteten. Erwartung und Wert erklärten 14% der Varianz in der Testleistung. Allerdings stellte nur der Interessenaspekt einen signifikanten Prädiktor von Testleistung dar. In einer anderen Studie (Freund & Holling, 2011), die denselben Motivationsfragebogen einsetzte, waren Erfolgserwartung und Interesse signifikante Prädiktoren der Leistung in einem Intelligenztest auch nach Kontrolle allgemeiner kognitiver Fähigkeit. Testteilnahmemotivation und kognitive Fähigkeit erklärten zusammen fast ein Drittel der Varianz der Testergebnisse.

Eine aktuelle internationale Studie erkannte (Knekta & Eklöf, im Druck) ebenfalls die Notwendigkeit, Erwartung, Wert und Anstrengung für die Modellierung der Testteilnahmotivation zu erheben. In ihrer Studie erfassten die Autorinnen daher Erfolgserwartung, Wichtigkeit, Interesse sowie Testangst als Prädiktoren für die Anstrengungsbereitschaft. Dabei sagten die wahrgenommene Wichtigkeit des Tests und die Erfolgserwartungen die investierte Anstrengung vorher. Bezüglich des Zusammenhangs mit Testleistung in Biologie, Chemie oder Physik wurden Anstrengungsbereitschaft und die Note (in Biologie, Chemie oder Physik) als Prädiktoren für die korrespondierende Testleistung verwendet. Dabei zeigte die investierte Anstrengung einen signifikanten Zusammenhang mit der Testleistung über die Note hinaus. Allerdings gab es keine weiteren Analysen zum Zusammenhang von Erwartung, Wert und Testleistung.

*Domänenspezifische und situationsspezifische Motivation*

Die Erwartung-Wert-Theorie wurde im Bereich der domänenspezifischen Leistungsmotivation entwickelt und validiert (Wigfield & Eccles, 2000), so dass der Einsatz dieses Modells ohne Anpassung für die Erfassung der situativen Testteilnahmemotivation problematisch ist (Eklöf, 2010a). Ob die theoretische (domänenspezifische) Konzeption auch auf den Kontext der (situativen) Testteilnahmemotivation übertragbar ist, ist noch zu erforschen. In vielen Studien werden domänenspezifische Kompetenzüberzeugungen zur Operationalisierung der Erwartungskomponente eingesetzt; aber der Wert des Tests wird situationsspezifisch erfasst (Barry & Finney, im Druck; Eklöf, 2007). Das bedeutet, es wird zwar die wahrgenommene Wichtigkeit des zu bearbeitenden Tests erfragt, jedoch die allgemeine Erfolgserwartung in Mathematik (und nicht die Erfolgserwartung für den zu bearbeitenden Test). Die theoretisch distinkten domänenspezifischen Kompetenzüberzeugungen und situationsspezifischen Erfolgserwartungen hängen zwar empirisch stark miteinander zusammen (Wigfield & Eccles, 2000); jedoch ist es vor allem im Kontext der Testteilnahmemotivation notwendig, zwischen situationsspezifischen Erfolgserwartungen („Kann ich *diesen* Mathematiktest schaffen?“) und domänenspezifischen Kompetenzüberzeugungen („Bin ich gut in Mathematik?“) zu unterscheiden. Die Unterscheidung kann erneut durch das Beispiel aus Abschnitt 2.4.1 illustriert werden: Eine Schülerin mag zwar das Fach Mathematik und ist sehr gut darin; sie bearbeitet einen Low-Stakes-Assessment in Mathematik trotzdem unmotiviert, da sie keine positiven Konsequenzen zu erhoffen oder negativen Konsequenzen zu befürchten hat.

Weiterhin ist unklar, ob die Erfolgserwartungen einen Zusammenhang mit Testleistung aufweisen, wenn für die domänenspezifischen Kompetenzüberzeugungen kontrolliert wird. Lediglich Eklöf (2007, 2008) untersuchte die Testteilnahmemotivation schwedischer Achtklässlerinnen und Achtklässler in TIMSS 2003 und ihre Analysen ergaben, dass nach Kontrolle des Selbstkonzepts in Mathematik und des wahrgenommenen Wertes in Mathematik die Anstrengungsbereitschaft nicht mehr signifikant (Eklöf, 2007) oder nur sehr schwach (Eklöf, 2008) mit Mathematikleistung zusammenhängt. Auch hier wurde, ähnlich zu den meisten Studien, Testteilnahmemotivation durch Fragen zur Anstrengungsbereitschaft und zur Wichtigkeit des Tests gemessen.

*Veränderung von Testteilnahmemotivation während einer Testsitzung*

Neuere Studien beschäftigen sich mit dem Verlauf von Testteilnahmemotivation innerhalb einer mehrstündigen Testsitzung (Barry & Finney, im Druck; Barry, Horst, Finney, Brown & Kopp, 2010; Horst, 2010). Dies ist relevant, da die meisten Low-Stakes-Assessments in der Sekundarstufe I oft mehr Zeit benötigen (i.d.R. 120 Minuten für den Leistungstest plus einer Pause nach den ersten 60 Minuten) als eine reguläre Klassenarbeit, die meist eine Schulstunde (d. h. 45 Minuten) dauert. Daher kann eine ungewohnt lange Testsitzung möglicherweise zu Ermüdungseffekten und einer Abnahme in der Testteilnahmemotivation bei den Schülerinnen und Schülern führen sowie zu einer niedrigen Testleistung. Dabei kann Ermüdung als das Ausmaß definiert werden, zu dem die Teilnehmenden während der Testsitzung ermüden oder gelangweilt werden (Barry, 2010). Während eines mehrstündigen High-Stakes-Tests berichteten beispielsweise die Teilnehmenden, die eine Abnahme in der Anstrengungsbereitschaft angaben, eine höhere Ausprägung der Ermüdung als die Teilnehmenden, die eine konstante oder ansteigende Anstrengungsbereitschaft über die Testsitzung angaben (Ackerman & Kanfer, 2009). Daher ist es denkbar, dass gerade in Low-Stakes-Tests aufgrund der fehlenden Konsequenzen Ermüdung die Testteilnahmemotivation in der Form beeinflusst, dass je länger ermüdete Teilnehmenden einen Leistungstest bearbeiten müssen, desto größer die Abnahme der Motivation ist. Dabei ist Ermüdung nur eine Erscheinungsform der motivationalen Abnahme, die sich in einer sinkenden Anstrengungsbereitschaft manifestieren kann. Da Testteilnahmemotivation als dreidimensionales Konstrukt (Erwartung, Wert und Anstrengung) konzipiert ist, kann auch eine Abnahme der Erfolgserwartungen während eines Leistungstests unabhängig von Ermüdungserscheinungen zu einer sinkenden Motivation führen.

Einige Studien untersuchten den Verlauf der Anstrengungsbereitschaft und der wahrgenommenen Wichtigkeit des Tests innerhalb einer dreistündigen Testsitzung, in der Studierende einen Leistungstests und vier Einstellungsfragebögen (im Original nichtkognitive Tests genannt, wie z. B. Einstellung zur Universität) bearbeiteten (Barry & Finney, im Druck; Barry et al., 2010; Horst, 2010). Bemerkenswerterweise fanden sie insgesamt keine Hinweise für Ermüdungseffekte innerhalb der gesamten Testsitzung. Allerdings variierte die Anstrengungsbereitschaft in Abhängigkeit des Assessmenttyps. Die Studierenden waren mehr gewillt, sich bei der Beantwortung der leichteren Einstellungstests anzustrengen als bei der Beantwortung des Leistungstests, obwohl sie letzteren als wichtiger wahrnahmen (Barry & Finney, im Druck; Barry et al., 2010). Für den Leistungstest

korrelierten Anstrengungsbereitschaft und Wichtigkeit moderat miteinander, allerdings gab es keinen Zusammenhang zwischen der Veränderung in diesen beiden Konstrukten (Barry & Finney, im Druck). Bisher untersuchte nur Horst (2010) den Verlauf der Anstrengungsbereitschaft innerhalb eines *Paper-and-Pencil-Leistungstests*. In ihrer Stichprobe gab es eine leichte Annahme der Anstrengungsbereitschaft und einen stabilen Verlauf der wahrgenommenen Wichtigkeit. Unerforscht ist bisher, ob der Trend einer abnehmenden Anstrengungsbereitschaft auch in anderen Studien nachzuweisen ist. Ebenfalls zeigen Studien, die Itempositionseffekte in Leistungsstudien untersuchen, dass Aufgaben umso schwieriger werden, je weiter hinten sie im Testheft erscheinen, was auf absinkende Anstrengungsbereitschaft zurückgeführt werden könnte (Cao & Stokes, 2008; Debeer & Janssen, 2013; Hohensinn, Kubinger, Reif, Schleicher & Khorramdel, 2011).

*Fazit*

Insgesamt verdeutlichen die meisten Studien, dass in Low-Stakes-Testsituationen die Teilnehmenden eine niedrigere Motivation berichten als in High-Stakes-Testsituationen und dass unmotiviertes Verhalten zu einer niedrigeren Testleistung führen kann (Sundre & Kitsantas, 2004; Wise & DeMars, 2005; Wolf & Smith, 1995). Der Forschungsstand zeigt aber auch auf, dass einige Sachverhalte bisher noch unklar sind. Auf drei davon soll hauptsächlich im Rahmen dieser Arbeit eingegangen werden. Erstens ist unerforscht, ob die situationsspezifische Testteilnahmemotivation überhaupt einen Zusammenhang mit Testleistung aufweist, wenn domänenspezifische Kompetenzüberzeugungen mit in die Analysen einbezogen werden (Eklöf, 2007, 2008). Wenn Testteilnahmemotivation keinen eigenen Beitrag zur Erklärung der Ergebnisse in Low-Stakes-Tests leistet unter Berücksichtigung domänenspezifischer Motivation, wäre eine weitere Erforschung des Zusammenhangs zwischen Testteilnahmemotivation und Leistung nicht mehr relevant.

Zweitens fehlt es im Kontext von Low-Stakes-Assessments an Untersuchungen, die, basierend auf der Erwartung-Wert-Theorie, alle drei Komponenten (sowohl Erwartung und Wert als auch Anstrengungsbereitschaft) bei der Erforschung von Testteilnahmemotivation berücksichtigen (Asseburg, 2011; Swerdzewski et al., 2011). Damit einhergeht die Diskussion, ob Anstrengungsbereitschaft den Effekt von Erwartung auf Leistung genauso vermittelt wie den Effekt von Wert auf Leistung (Cole et al., 2008). Der theoretische Erkenntnisgewinn beinhaltet insgesamt die Überprüfung, ob das Erwartung-Wert-Modell überhaupt als Grundlage geeignet ist, um Testteilnahmemotivation zu erfassen.

Drittens gibt es neue Studien, die sich mit dem Verlauf von Testteilnahmemotivation während der Bearbeitung eines Leistungstests beschäftigen. Diese Studien sind relevant, da beispielsweise Ermüdungseffekte zu einer sinkenden Anstrengungsbereitschaft während der für den Schulalltag unüblich langen Testsitzung führen können. Bislang wurde nicht die Veränderung in der Erwartungskomponente untersucht sowie die Beziehung zwischen der Veränderung in der Testteilnahmemotivation und der tatsächlich gezeigten Testleistung (Barry & Finney, im Druck; Barry et al., 2010; Horst, 2010).

Damit die aufgezeigten Sachverhalte untersucht werden können, wird im nächsten Abschnitt ein theoretisches Erwartung-Wert-Anstrengung-Modell der Testteilnahmemotivation aufgestellt, das auf der Erwartung-Wert-Theorie basiert und die Ergebnisse der aufgezeigten empirischen Studien berücksichtigt.

### 2.4.4 Ein Erwartung-Wert-Anstrengung-Modell der Testteilnahmemotivation

Wie in Abbildung 2.5 zu entnehmen ist, ist das originale Erwartung-Wert-Modell der Leistungsmotivation von Eccles et al. (2002) äußerst komplex, da es neben motivationalen Aspekten auch kognitive Prozesse sowie die soziale Umwelt der Individuen mit einbezieht. Um die situationsspezifische Testteilnahmemotivation in großangelegten Leistungsstudien zu modellieren, sind nicht alle Aspekte erforderlich, so dass für diese Arbeit auf ein Teilmodell fokussiert wird. So werden die soziale Welt und die kognitiven Prozesse aus der linken Hälfte in Abbildung 2.5 in dieser Arbeit ausgeklammert und der Fokus auf die Hauptelemente des Prozesses der Testteilnahmemotivation gelegt.

Der bisherige Forschungstand macht deutlich, dass es an einem spezifischen Modell für Testteilnahmemotivation mangelt. Zwar hat Asseburg (2011) in ihrer Dissertation ein an die Erwartung-Wert-Theorie angepasstes Modell zur Motivation der Testbearbeitung für ihre Studie aufgestellt, aber die Anstrengungsbereitschaft nicht berücksichtigt. Anstrengungsbereitschaft konstituiert einen zentralen Aspekt in der Forschung zur Testteilnahmemotivation in Low-Stakes-Assessments, wie in der Definition von Testteilnahmemotivation deutlich wird. Jedoch wird diese im originalen Modell von Eccles und Wigfield unterschiedlich lokalisiert: Zum einen wird sie als Ergebnis von Erwartung und Wert konzeptualisiert (Wigfield & Eccles, 2000), zum anderen als ein Teil des Kostenaspekts (Eccles & Wigfield, 2002). In der Arbeit von Wolf, Smith und Birnbaum (1995) bildet die Anstrengungsbereitschaft, neben der Erfolgswahrscheinlichkeit, einen Teil der Erwartungskomponente. Zunächst unabhängig von der Verortung im Modell, ist es

unumstritten, dass ohne Beachtung dieses zentralen Konstruktes ein Modell für Testteilnahmemotivation unvollständig wäre (Baumert & Demmrich, 2001; Sundre, 2007; Wise & Smith, 2011). Daher wird im Folgenden ein neues, an Testteilnahmemotivation angepasstes Erwartung-Wert-Modell entwickelt, das in Abbildung 2.7 illustriert ist.

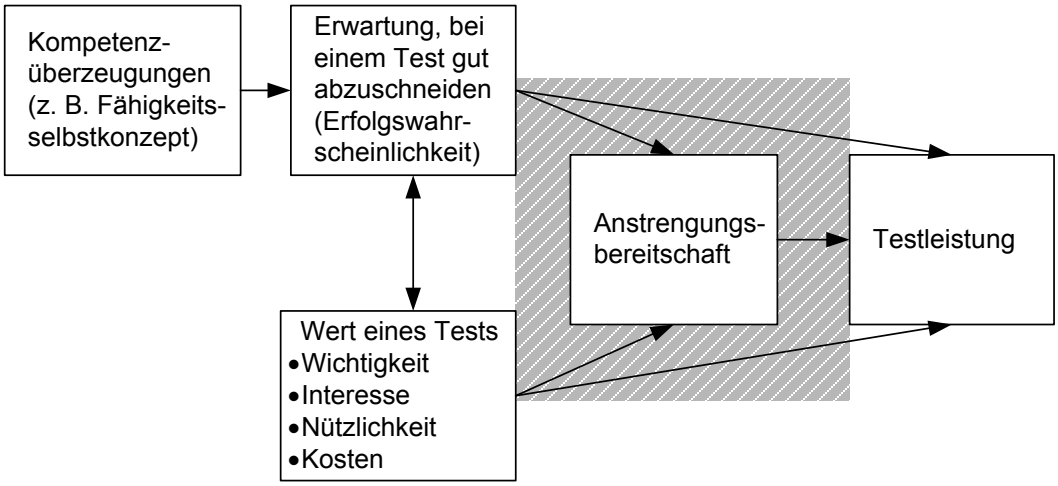



*Abbildung 2.7.* Erwartung-Wert-Anstrengung-Modell der Testteilnahmemotivation (grau hinterlegt ist die konzeptuelle Veränderung zum originalen Erwartung-Wert-Modell).

In einem an Testteilnahmemotivation adaptierten Erwartung-Wert-Modell sind die drei zentralen Komponenten die Erfolgserwartung, der wahrgenommene Wert eines Tests und die Anstrengungsbereitschaft. Die Erwartungskomponente umschreibt die individuelle Erfolgswahrscheinlichkeit einer Testperson, in einem Leistungstest erfolgreich abzuschneiden („Bin ich fähig, *diesen Test* zu bewältigen?“), und ist damit in ähnlicher Weise konzipiert wie im originalen Erwartung-Wert-Modell. Theoretisch wird die Erfolgserwartung von den domänenspezifischen Kompetenzüberzeugungen beeinflusst, die im originalen Modell auch einen sehr schwachen Einfluss auf die Wertkomponente zeigen. Wie in der Arbeit von Asseburg (2011) wird jedoch nur ein Zusammenhang zwischen Kompetenzüberzeugungen und Erwartung modelliert. Konzeptuell ist es sinnvoll, dass domänenspezifische Kompetenzüberzeugungen eher die Erfolgserwartungen, in einem Low-Stakes-Test gut abzuschneiden, beeinflussen als den wahrgenommenen Wert des Tests. An dieser Stelle ist wichtig, zu erwähnen, dass sich domänenspezifische Kompetenzüberzeugungen nicht von Motivation trennen lassen. Allerdings ist eine Differenzierung zwischen situationsspezifischer Motivation, die in dieser Arbeit im Fokus steht, und domänenspezifischer Motivation aufgrund der Definition von Testteilnahmemotivation notwendig (s. auch Fragestellung von Studie I).

Die Wertkomponente besteht aus den oben genannten vier Aspekten und beinhaltet damit, wie wichtig, interessant und nützlich eine erfolgreiche Testbearbeitung für eine Person ist sowie die Kosten der Testbearbeitung („Warum sollte ich *diesen Test* bearbeiten?“). Auch damit ähnelt die Konzeption der Wertaspekte der Konzeption im Originalmodell, nur dass der Fokus nicht auf den Wert einer Aufgabe gesetzt wird, sondern auf den Wert eines Tests. Der Wichtigkeitsaspekt bezieht sich dementsprechend darauf, wie bedeutsam die Testbearbeitung für die Person ist. Dass dabei, analog zum Originalmodell, zentrale Aspekte des Selbstschemas bestätigt werden, ist eher unwahrscheinlich, da in dem angepassten Modell auf den individuell wahrgenommen Wert einer guten Testbearbeitung fokussiert wird, die ohne Folgen bleibt. Eine Ausnahme bilden Schülerinnen und Schüler, die sich selbst als gute und interessierte Personen wahrnehmen. Hier wird unter Wichtigkeit auch gefasst, ob die Testteilnehmenden die Situation als eine Leistungssituation wahrnehmen. Beispielsweise kann ein Leistungstest für einen Schüler dann an Bedeutung gewinnen, wenn er sich durch ihn herausgefordert fühlt und er ihn erfolgreich bearbeiten möchte (Vollmeyer & Rheinberg, 2006). Interesse steht in Verbindung mit der wahrgenommenen Freude während der Testbearbeitung. Wenn die Testbearbeitung den Teilnehmenden Spaß bereitet, zum Beispiel durch ansprechend gestaltete Aufgaben, ist es wahrscheinlich, dass sie sich anstrengen und den Test bis zum Schluss engagiert bearbeiten. Der Nützlichkeitsaspekt ist im Kontext von Testteilnahmemotivation kompliziert, da die Passung einer Testbearbeitung mit zukünftigen Zielen schwer vorstellbar ist, wenn selbst ein erfolgreiches Abschneiden keine positiven Konsequenzen nach sich zieht. Ob den Testteilnehmenden der Nutzen einer motivierten Beantwortung der Testaufgaben für die Evaluation des nationalen Bildungssystems bewusst ist, bleibt unklar. Daher scheint es, dass der Nützlichkeitswert per se äußerst niedrig ist (Barry & Finney, im Druck), was möglicherweise erklärt, warum dieser Aspekt so selten erhoben wird. Die Kostenkomponente wurde im originalen Erwartung-Wert-Modell bisher am wenigsten untersucht (Wigfield & Eccles, 2000). Im situativen Kontext der Testteilnahmemotivation beinhalten die Kosten die negativen Aspekte der Testbearbeitung. Bei groß angelegten Leistungsstudien ist es sinnvoll, hier die Testangst oder auch Sorgen und Ablenkungen vom Test zu verorten. Bei der Bearbeitung von Low-Stakes-Assessments können generell testängstliche Schülerinnen und Schüler sich vor einem potentiellen Misserfolg fürchten (Eklöf & Nyroos, 2013; Wolf et al., 1995).

Die schon oft erwähnte Anstrengungsbereitschaft konstituiert einen zentralen Aspekt in der Forschung zur Testteilnahmemotivation in Low-Stakes-Assessments. Allerdings wird sie, wie oben skizziert, in unterschiedlichen Arbeiten an verschiedenen Stellen im Modell verortet. Wie in Abbildung 2.7 verdeutlicht, wird für diese Arbeit Anstrengungsbereitschaft als Ergebnis von Erwartung und Wert angesehen, da es theoretisch (Wigfield & Eccles, 2000) und empirisch (Cole et al., 2008) plausibel ist, dass beide Komponenten in eine hohe Anstrengungsbereitschaft münden. Diese Konzeptualisierung der Anstrengungsbereitschaft als Ergebnis von Erwartung und Wert ist die hervorstechendste Anpassung des originalen Erwartung-Wert-Modells an die Spezifika der Testteilnahmemotivation. Das heißt, wenn Schülerinnen und Schüler bezüglich ihrer Erfolgswahrscheinlichkeit positiv eingestellt sind und den Test wertschätzen, sind sie bereit, Anstrengung zu investieren. Und Schülerinnen und Schüler, die eine hohe Anstrengungsbereitschaft aufweisen, sollten eine höhere Testleistung zeigen. Ebenfalls haben motivationale Dispositionen einen Einfluss auf die zwei Komponenten. So nimmt beispielsweise das fähigkeitsbezogene Selbstkonzept Einfluss auf die Erwartungskomponente. Bisher untersuchten Studien (Eklöf, 2007, 2008) lediglich den Einfluss der Testteilnahmemotivation unter Berücksichtigung domänenspezifischer Kompetenzüberzeugungen, ohne die im Erwartung-Wert-Modell postulierte Wirkrichtung zu prüfen.

Zusammengefasst sind die wichtigsten Komponenten im aufgestellten Erwartung-Wert-Anstrengung-Modell der Testteilnahmemotivation die Erfolgserwartung, einen Test bewältigen zu können, der wahrgenommene Wert, der einer erfolgreichen Testbearbeitung zugeschrieben wird, sowie die Anstrengungsbereitschaft. Die Erwartungskomponente wird wiederum von domänenspezifischen Kompetenzüberzeugungen beeinflusst, so dass neben situationsspezifischen, instabilen Merkmalen auch stabile Personenmerkmale berücksichtigt werden, wie im Grundmodell der Motivationspsychologie gefordert. Die Anstrengungsbereitschaft ist als Ergebnis von Erwartung und Wert konzeptualisiert und besitzt einen Einfluss auf die Testleistung. Dabei sind sowohl direkte Effekte von Erwartung und Wert auf Leistung modelliert, als auch indirekte Effekte von Erwartung und Wert auf Leistung via Anstrengungsbereitschaft. Dieses Modell bildet die theoretische Grundlage dieser Arbeit, die überprüfen wird, ob das postulierte Modell sich in der Empirie als geeignet erweist.

# 3

# Ziel und Fragestellungen

# 3 Ziel und Fragestellungen

Wie der vorliegende Forschungsstand zeigt, kann fehlende Motivation auf Seiten der Testteilnehmenden eine Gefahr für die valide Interpretation der Testergebnisse sein, die mithilfe von Low-Stakes-Tests angestrebt wird. Bei der Verwendung der Ergebnisse zur Evaluation der Qualität von Bildungssystemen wird unterstellt, dass die Testteilnehmenden während der Testung ihr Bestes geben und diese Anstrengung auch bis zur letzten Aufgabe aufrechterhalten (Thelk et al., 2009). Es scheint allerdings, dass sich während der Testbearbeitung bei den Teilnehmenden die investierte Anstrengungsbereitschaft verringert, auch wenn sie den Test als wichtig erachten (Horst, 2010). Dies wird außerdem von Studien zu Itempositionseffekten untermauert, die aufzeigen, dass Aufgaben umso schwieriger sind, je weiter hinten sie im Testheft auftauchen (Debeer & Janssen, 2013). Daher sind Kenntnisse über die motivationalen Prozesse während der Testsitzung und über den Einfluss der Testteilnahmemotivation auf die Testleistung unabdingbar.

Forschung in diesem Kontext basiert häufig auf der Erwartung-Wert-Modell der Leistungsmotivation. Allerdings wurde kein theoretisches Modell entwickelt, dass sowohl an die Besonderheiten der Testteilnahmemotivation angepasst ist als auch alle drei Komponenten (Erwartung, Wert und Anstrengung) berücksichtigt. Die Erforschung der Testteilnahmemotivation und deren Beziehung zur Testleistung ist theoretisch bedeutsam, um die in Abschnitt 2.4.4 vorgenommenen Anpassungen an das Erwartung-Wert-Modell empirisch zu prüfen und so ein geeignetes Modell der Testteilnahmemotivation für den Large-Scale-Assessment-Kontext zur Verfügung zu stellen. Damit trägt die Dissertation zu einem fortlaufenden akademischen Diskurs bei. Das Wissen darüber, welche motivationalen Prozesse in einer Schülerin oder einem Schüler während der Bearbeitung eines Leistungstests ohne persönliche Konsequenzen ablaufen, kann auch für die Konstruktion von Large-Scale-Assessments Verwendung finden. Die vorliegende Arbeit untersucht daher die Relationen zwischen den verschiedenen Komponenten der Testteilnahmemotivation sowie den Zusammenhang zwischen Testteilnahmemotivation und Testleistung. Dabei wird jeweils auf einzelne Aspekte des in Abschnitt 2.4.4 entwickelten Erwartung-Wert-Anstrengung-Modells der Testteilnahmemotivation fokussiert.

Die vorliegenden Studien weisen Gemeinsamkeiten und Unterschiede auf. Allen Studien ist gemein, dass Schülerinnen und Schüler der Sekundarstufe I die Stichprobe bilden, da es besonders kompliziert zu sein scheint, Jugendliche für die Teilnahme an Low-

Stakes-Assessments zu begeistern. Auch Studien zum Erwartung-Wert-Modell im domänenspezifischen Bereich zeigten im Verlauf der Sekundarstufe I eine generelle Abnahme fähigkeitsbezogener Überzeugungen und des subjektiven Wertes bestimmter akademischer Aufgaben sowie von Leistung und Anstrengung generell, die die Jugendlichen zeigen (Wigfield & Cambria, 2010; Wigfield & Eccles, 2000). Daher scheint ein geringes Interesse an einem Test plausibel, der ohnehin keine negativen oder positiven Folgen für die Jugendlichen enthält. Dies untermauert die Notwendigkeit der Berücksichtigung motivationaler Effekte bei der Interpretation von Testergebnissen aus Low-Stakes-Tests vor allem in der Sekundarstufe I. Außerdem konzentrieren sich die drei Studien auf die Testleistung in der Domäne Mathematik. Zum einen wurde das Erwartung-Wert-Modell intensiv für dieses Fach empirisch erforscht (Wigfield & Eccles, 2000). Zum anderen gehört das Fach Mathematik, neben Deutsch, den Naturwissenschaften und der Fremdsprache Englisch, zu einem der vier Kernbereiche schulischer Bildung, in dem der Erwerb gesicherter Kompetenzen für eine erfolgreiche Teilhabe an der Gesellschaft und am Leben unabdingbar ist (Köller & Baumert, 2012). Die Aussagen über die mathematische Kompetenz von Schülerinnen und Schülern mithilfe von Large-Scale-Assessments sollten daher nicht mit Aussagen über motivationale Effekte konfundiert sein. Die Unterschiede bestehen zum einen in dem Erhebungsjahr der analysierten Daten (Studie I aus dem Jahr 2000, Studie II und III aus dem Jahr 2012). Zum anderen wurden, je nach Fragestellung, unterschiedliche Teilaspekte des in Abschnitt 2.4.4 entworfenen Erwartung-Wert-Anstrengung-Modells erforscht.

*Studie I*

Um die Beziehung zwischen den verschiedenen Komponenten der Testteilnahme und Testleistung untersuchen zu können, muss zunächst geklärt sein, ob die situationsspezifischen Motivationskomponenten überhaupt mit Testleistung in Large-Scale-Assessments zusammenhängen, wenn domänenspezifische motivationale Merkmale wie das Selbstkonzept berücksichtigt werden. Dass domänenspezifische Kompetenzüberzeugungen einen Zusammenhang mit der korrespondierenden Leistung aufweisen, wurde vielfach belegt (Chen, Yeh, Hwang & Lin, 2013; Jansen et al., 2013). Die erste Fragestellung der Studie I überprüft, ob über die generelle domänenspezifische Selbsteinschätzung der Kompetenz hinaus die situationsspezifische Motivation, wie zum Beispiel Anstrengungsbereitschaft, zusätzlich Unterschiede in der Testleistung erklären kann. Die zwei bekannten Studien (Eklöf, 2007, 2008) zu dieser Frage kamen zu widersprüchlichen Ergebnissen, so dass

keine eindeutigen Antworten vorliegen und keine Annahme formuliert werden kann (s. Abschnitt 2.4.3). Daher fokussiert die erste Studie zum einen, wie in Abbildung 3.1 zu sehen ist, auf den Zusammenhang zwischen Testleistung und Testteilnahmemotivation (hier die Wertkomponente und Anstrengungsbereitschaft) unter Berücksichtigung domänenspezifischer Kompetenzüberzeugungen. Aufgrund der Fragestellung und da die Erfolgserwartungen nicht erfasst wurden, wird ein direkter Zusammenhang der Kompetenzüberzeugungen mit Testleistung modelliert.

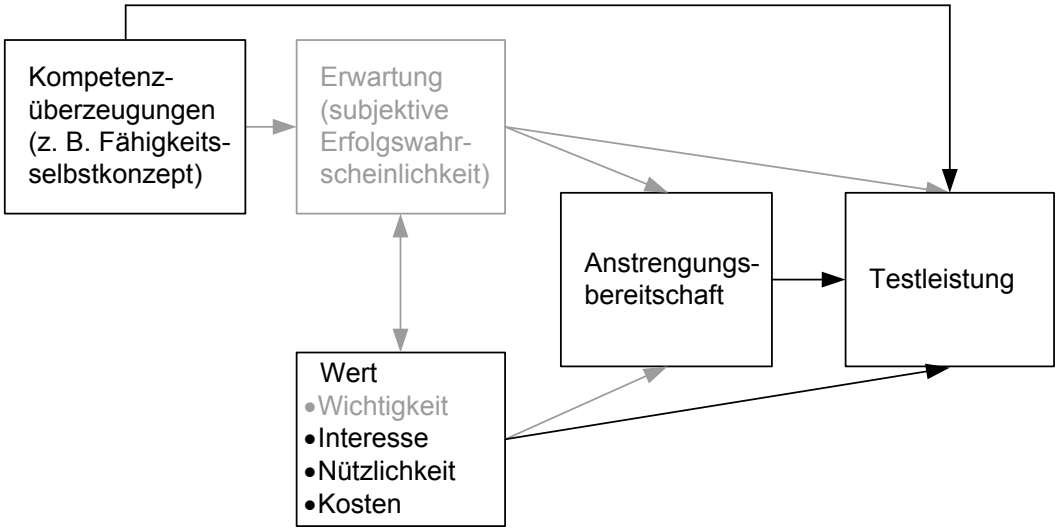



*Abbildung 3.1*. Verortung der Fragestellung 1 von Studie I: Prädiktion von Leistung durch domänenspezifische Kompetenzüberzeugungen, Wert und Anstrengung.

Zum anderen wurde der Zusammenhang zwischen der Wertkomponente und Anstrengungsbereitschaft untersucht. Wie in Abschnitt 2.4.3 beschrieben, wird in den meisten Studien Testteilnahmemotivation über das Konstrukt der Anstrengungsbereitschaft operationalisiert (Wise & DeMars, 2005). Allerdings wird der theoretisch und empirisch postulierte Zusammenhang zwischen Anstrengung und der Wertkomponente (Cole et al., 2008) eher selten modelliert. Daher wird in einem zweiten Schritt analysiert, ob die erfassten situationsspezifischen Wertaspekte Prädiktoren der investierten Anstrengungsbereitschaft darstellen (s. Abbildung 3.2). Es wird angenommen, dass vor allem der Interessenaspekt (d. h. die Freude während der Testbearbeitung) sowie die Kosten der Wertkomponente (z. B. Testangst) die Unterschiede in der investierten Anstrengungsbereitschaft erklären können, da der Nützlichkeitsaspekt per se in Low-Stakes-Assessments niedrig ist. Da die Erfolgserwartungen nicht erhoben wurden, kann deren Zusammenhang mit Anstrengungsbereitschaft nicht untersucht werden.

Ein weiterer Aspekt ist die Untersuchung der Unterschiede von Testteilnahmemotivation nach Schulart. Bisher gibt es nur eine nationale Studie im Bereich Low-Stakes-Assessments, die die Testteilnahmemotivation von Teilnehmenden am Gymnasium und Teilnehmenden an der Hauptschule erforschte und herausfand, dass Jugendliche an Hauptschulen eine ungünstigere Testteilnahmemotivation berichteten als Jugendliche an Gymnasien (Baumert & Demmrich, 2001). Daher werden zusätzlich beide Fragestellungen dieser Studie differenziert betrachtet, ob sich die Ergebnisse der zwei vorgestellten Modelle (Abbildung 3.1 und 3.2) für Teilnehmende an Gymnasium und für Teilnehmende an nicht-gymnasialen Schularten unterscheiden. Konkrete Annahmen zu schulartspezifischen Unterschieden bezüglich der beiden Fragestellungen werden nicht aufgestellt, da es an theoretischen und empirischen Vorinformationen mangelt.

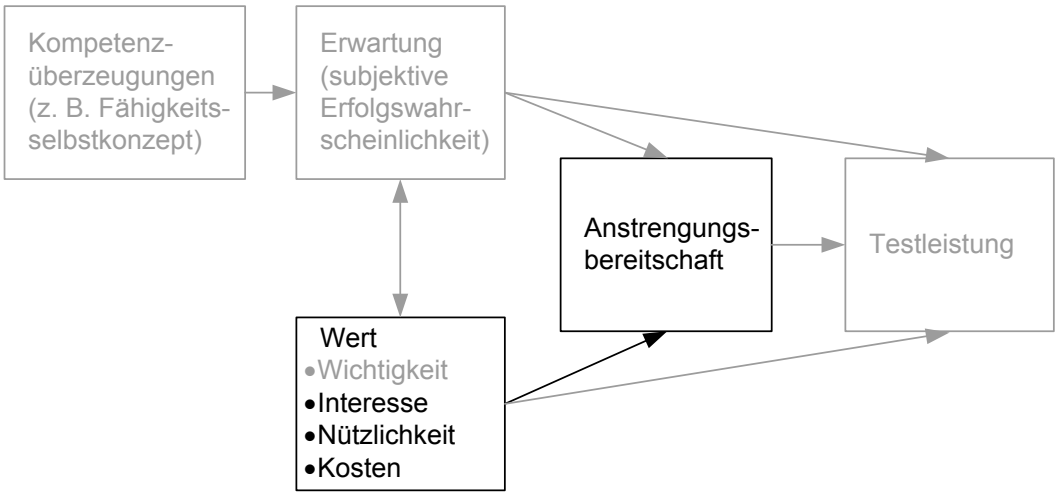



*Abbildung 3.2*. Verortung der Fragestellung 2 von Studie I: Prädiktion von Anstrengungsbereitschaft durch den Wert.

Die Datengrundlage für die erste Studie bildet die erste PISA-Erhebung aus dem Jahr 2000. Als Testteilnahmemotivation wurden die Skalen investierte Anstrengungsbereitschaft, Testattraktivität (entspricht dem Interesse am Test) sowie Nützlichkeit des Tests erhoben. Darüber hinaus wurde die emotionale Befindlichkeit der Schülerinnen und Schüler sowie deren Sorgen bezüglich der eigenen Fähigkeiten und Ablenkung vom Test (d. h. inwiefern sich die Teilnehmenden gedanklich mit Test irrelevanten Dingen beschäftigt haben) erfasst. Die zwei zuletzt genannten Skalen stellen sogenannte lösungsirrelevante Kognitionen als Komponenten der Testangst dar und können dem Kostenaspekt zugeordnet werden. Der emotionale Zustand hat kein direktes theoretisches Gegenstück in dem Modell, kann allerdings als Pendant zu den berichteten Sorgen und Ablenkung

angesehen werden und damit auch den Kosten zugeordnet werden. Als domänenspezifische Kompetenzüberzeugungen wurde das Selbstkonzept in Mathematik von den Schülerinnen und Schülern erfragt. Für diese Studie lagen keine Daten zu den situationsspezifischen Erfolgserwartungen und der wahrgenommenen Wichtigkeit des Tests vor.

Zur Beantwortung der Fragestellungen wurden Regressionsanalysen durchgeführt. Zur Überprüfung des ersten Modells wurde die Testleistung in Mathematik mit den verschiedenen Skalen der Wertkomponente und der Anstrengungsbereitschaft vorhergesagt. Selbstkonzept in Mathematik wurde als Kovariate mit in das Modell aufgenommen. Als Personenschätzer der Fähigkeit in Mathematik dienten fünf *Plausible Values*. Für die Analyse des zweiten Modells wurde Anstrengungsbereitschaft mit den Skalen der Wertkomponente vorhergesagt. Beide Analysen wurden zusätzlich für Teilnehmende an Gymnasien und Teilnehmenden an nicht-gymnasialen Schularten separat durchgeführt.

*Studie II*

Wie der Forschungsstand in 2.4.3 aufzeigte, wurde in den meisten Studien mindestens eine Komponente des Erwartung-Wert-Modells vernachlässigt. Vor allem internationale Studien konzentrierten sich lediglich auf die Wertkomponente und Anstrengungsbereitschaft (Cole et al., 2008; Eklöf & Nyroos, 2013; Eklöf et al., 2014; Swerdzewski et al., 2011; Wolf & Smith, 1995) und nationale Studien ausschließlich auf die Erwartung- und Wertkomponente (Asseburg, 2011; Freund & Holling, 2011; Freund et al., 2011). Durch die Reanalyse bereits vorhandener Daten der PISA-Studie konnte die Erwartungskomponente nicht mit in die Berechnungen für die erste Studie dieser Dissertation einbezogen werden. Daher ist die Zielstellung der zweiten Studie die Untersuchung des komplexen Zusammenspiels von Erwartung, Wert, Anstrengung und Leistung, um zu überprüfen, ob sich das theoretisch postulierte Erwartung-Wert-Anstrengung-Modell empirisch bestätigen lässt. Vor allem die Untersuchung der bisher vernachlässigten Zusammenhänge zwischen Erwartungskomponente und Anstrengungsbereitschaft sowie Leistung trägt zum theoretischen Erkenntnisgewinn bei.

Als erster Schritt wird getestet, ob Unterschiede in der Anstrengungsbereitschaft sowohl durch die Wert- als auch durch die Erwartungskomponente erklärt werden können, wie es in der Studie von Knekta und Eklöf (im Druck) der Fall war. Die konkreten Annahmen folgen in der Beschreibung des finalen Modells. Anschließend werden zur

Erklärung der Unterschiede in der Testleistung alle drei Komponenten des Erwartung-Wert-Anstrengung-Modells mit in die Analyse einbezogen. Das zu testende finale Modell ist in Abbildung 3.3 dargestellt. Neben der Prädiktion von Testleistung durch die drei Komponenten Erwartung, Wert und Anstrengung wird ebenfalls geprüft, ob die Zusammenhänge zwischen Leistung und Erwartung beziehungsweise Leistung und Wert über Anstrengungsbereitschaft vermittelt werden. Die Prüfung der Mediation der Effekte der Erwartungs- und Wertkomponente auf Testleistung via Anstrengungsbereitschaft ist ein weiterer Zugewinn dieser Studie. Aufgrund der bisherigen Forschung wird angenommen, dass die Erwartungskomponente und die Aspekte der Wertkomponente die Anstrengungsbereitschaft vorhersagen (Abdelfattah, 2010; Barry & Finney, im Druck; Cole et al., 2008; Eklöf & Nyroos, 2013; Knekta & Eklöf, im Druck; Zilberberg et al., 2014). Außerdem wird ein indirekter Effekt der Wertkomponente (Cole et al., 2008; Zilberberg et al., 2014) auf Testleistung erwartet sowie ein direkter Effekt (Wigfield & Eccles, 2000). Aufgrund der Theorie wird auch ein Zusammenhang zwischen der Erwartungskomponente und Anstrengungsbereitschaft vermutet (Wigfield & Eccles, 2000). Darüber hinaus kann von einem direkten Zusammenhang zwischen der Erwartungskomponente und der Testleistung sowie der Anstrengungsbereitschaft und der Testleistung ausgegangen werden (Asseburg, 2011; Knekta & Eklöf, im Druck).

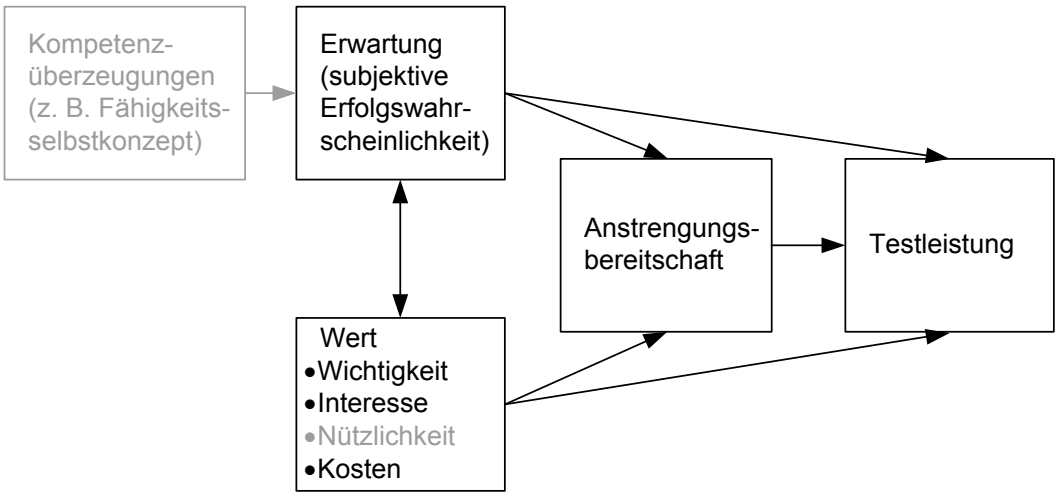



*Abbildung 3.3*. Verortung des finalen Modells der Studie II: Prädiktion von Leistung durch Erwartung, Wert und Anstrengungsbereitschaft.

Neben dem komplexen Beziehungsgeflecht der drei motivationalen Konstrukte untereinander und deren Beziehung zur Testleistung wird außerdem untersucht, ob der Zeitpunkt der Erfassung der Testteilnahmemotivation zu Veränderungen im Modell führt.

Konkret wird analysiert, ob sich der Zusammenhang zwischen Testteilnahmemotivation und Leistung verändert, je nachdem ob die initiale Testteilnahmemotivation vor dem Test als Prädiktor verwendet wird oder die Testteilnahmemotivation, die die Schülerinnen und Schüler nach der Bearbeitung der Testaufgaben berichten. In den meisten Studien ist es üblich, die Testteilnahmemotivation nach dem Leistungstest zu erfragen. Allerdings gibt es wenige Studien (Freund & Holling, 2011; Freund et al., 2011), die motivationalen Konstrukte erheben, bevor die Testteilnehmenden mit der Bearbeitung des Leistungstests begonnen haben, und zu divergenten Ergebnissen kommen. Ob die unterschiedlichen Ergebnisse der Studien, die die Motivation vor dem Test abfragen, auf den Zeitpunkt zurückgeführt werden können, soll mit diesen Analysen überprüft werden.

Die Datengrundlage der zweiten Studie bildet der Ländervergleich des Instituts zur Qualitätsentwicklung im Bildungswesen e. V. (IQB) aus dem Jahr 2012 (Pant et al., 2013). Damit liegen für die beschriebene Fragestellung aktuellere Daten vor als für die erste Studie. Den Teilnehmenden wurden im Ländervergleich Fragen zu ihrer Testteilnahmemotivation *vor* und *nach* dem Leistungstest gestellt. Konkret wurden die subjektive Erfolgswahrscheinlichkeit (Erwartung), Herausforderung (Wichtigkeit des Tests), Interesse am Test und Misserfolgsbefürchtungen (Kosten) erhoben. Dabei beschreibt das Konstrukt der Herausforderung, ob die Testteilnehmenden die Situation als Leistungssituation wahrnehmen sowie erfolgreich sein möchten und korrespondiert so mit der wahrgenommenen Wichtigkeit des Tests (Vollmeyer & Rheinberg, 2006). Daher wird im Folgenden auch der Begriff der Wichtigkeit verwendet. In dieser Teilstudie wurden keine Fragen zur empfundenen Nützlichkeit des Tests gestellt.

Zur Überprüfung der Eignung der eingesetzten Fragebögen wurden zunächst konfirmatorische Faktorenanalysen für die zwei Messzeitpunkte berechnet. Anschließend wurden latente Regressionsanalysen mit *Weighted Likelihood Estimates* als Personenschätzer der Fähigkeit in Mathematik getrennt für die Testteilnahmemotivation vor und nach dem Test durchgeführt. Neben den direkten Effekten wurden die indirekten Effekte der Erwartungs- und Wertkomponente auf die Testleistung via Anstrengungsbereitschaft berechnet.

*Studie III*

Die dritte und letzte Studie kombiniert die Erkenntnisse aus den ersten beiden Studien und testet diese in einem komplexen Erwartung-Wert-Anstrengung-Modell, das die Verände-

rung der Testteilnahmemotivation im Verlauf einer zweistündigen Testsitzung berücksichtigt. Die einzige Studie, die den Verlauf der Anstrengungsbereitschaft und Wichtigkeit des Tests innerhalb eines Leistungstests analysierte, ergab eine Abnahme in der Anstrengungsbereitschaft und einen stabilen Verlauf der wahrgenommenen Wichtigkeit (Horst, 2010). Allerdings wurde kein Zusammenhang zwischen der Veränderung der beiden Konstrukte oder zwischen der Veränderung in der Testteilnahmemotivation und der gezeigten Testleistung untersucht. Das *demands-capacity model of test-taking effort* (Wise & Smith, 2011) betont die dynamischen Prozesse bei der Konstituierung der Anstrengungsbereitschaft während einer Testsitzung und zeigt theoretisch die Beziehung zwischen der Veränderung der Erfolgserwartungen und der Wertaspekte sowie der Veränderung der Anstrengungsbereitschaft auf. Allerdings gibt es noch keine Studie, die diese theoretisch postulierten Zusammenhänge zwischen den Verläufen der motivationalen Konstrukte näher analysiert hat.

Daher wurde im IQB-Ländervergleich 2012 neben den zwei Messzeitpunkten vor und nach dem Test die Testteilnahmemotivation ebenfalls nach der ersten Hälfte der Tests erhoben. Die erste Fragestellung untersucht, ob sich die Erwartungs- und Wertkomponente sowie die Anstrengungsbereitschaft während der Testsitzung verändern. Aufgrund bisheriger Forschung (Cao & Stokes, 2008; Horst, 2010) wird eine Abnahme in der Anstrengungsbereitschaft und ein stabiler Verlauf der Wichtigkeit angenommen. Über den Verlauf der Erwartungskomponente kann aufgrund mangelnder Forschung keine Vorhersage getroffen werden. Wird allerdings davon ausgegangen, dass sich innerhalb eines Testheftes leichte und schwere Aufgaben abwechseln, wie es in der Assessmentpraxis üblich ist, ist ein stabiler Verlauf der Erwartung plausibel.

Als zweiter Schritt wird analysiert, ob die Verläufe der Erwartung, des Wertes und der Anstrengungsbereitschaft miteinander in Beziehung stehen. Obwohl die Studie von Barry und Finney (im Druck) keinen Zusammenhang zwischen der Veränderung im Wert und der Veränderung in der Anstrengungsbereitschaft fand, wird hier ein Zusammenhang erwartet. Die Annahme gründet sich auf den Fakt, dass in der amerikanischen Studie der Verlauf über verschiedene Testtypen (Leistungstests und Fragebögen) untersucht wurde und die vorliegende Studie den Verlauf innerhalb eines Leistungstests erforscht. Basierend auf dem *demands-capacity model of test-taking effort*, das die Veränderung in der Zuversicht, zukünftige Aufgaben zu lösen, mit einer Veränderung in der Anstrengungsbereit-

schaft in Beziehung setzt, wird außerdem davon ausgegangen, dass der Verlauf der Erwartung mit dem Verlauf der Anstrengungsbereitschaft einen Zusammenhang aufweist.

Das finale Modell überprüft, ob die Veränderungen in Erwartung, Wert und Anstrengungsbereitschaft einen Zusammenhang mit der gezeigten Testleistung zeigen. Abbildung 3.4 stellt das finale Modell dar und der Verlauf von Erwartung, Wert und Anstrengung ist durch die Hinzunahme des zweiten, grauen Kästchens mit der Uhr gekennzeichnet. Aufgrund fehlender vorheriger Forschung können keine Annahmen aufgestellt werden. Jedoch scheint es möglich, dass Teilnehmende mit einer abnehmenden Testteilnahmemotivation eine niedrigere Testleistung zeigen als Teilnehmende mit einem stabilen Verlauf. Insbesondere kann eine Abnahme in der Anstrengungsbereitschaft dazu führen, dass die Schülerinnen und Schüler möglicherweise nur die leichten Aufgaben beantworten oder den Test ganz abbrechen, was sich in einem insgesamt niedrigen Testergebnis widerspiegeln könnte. In dieser Analyse wird das Selbstkonzept in Mathematik als Prädiktor für die situationsspezifischen Erfolgserwartungen verwendet, da bekannt ist, dass domänenspezifische Kompetenzüberzeugungen sowohl theoretisch als auch empirisch einen Zusammenhang mit der Erwartungskomponente aufweisen (s. Abschnitt 2.3.2). Die Ergebnisse der zweiten Studien ergaben, dass der Wichtigkeitsaspekt der stärkste Prädiktor der Anstrengungsbereitschaft war und einen indirekten Effekt mit der Testleistung aufwies. Daher wird für die Wertkomponente nur die Veränderung des Wichtigkeitsaspekts modelliert.

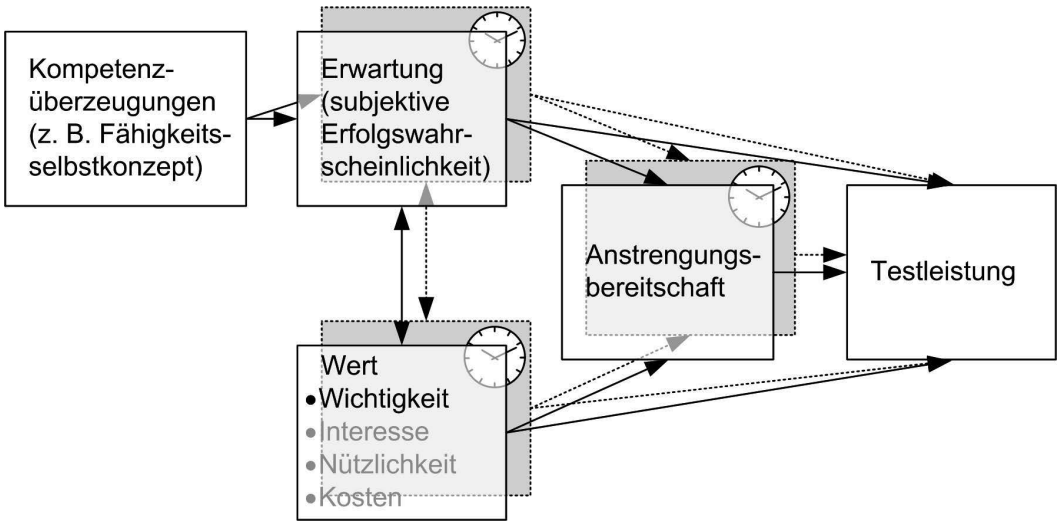



*Abbildung 3.4*. Verortung des finalen Modells von Studie III: Prädiktion von Testleistung durch Veränderung in der Erwartungs-, Wertkomponente und Anstrengungsbereitschaft.

Unter Anwendung von *second-order latent growth curve modeling* kann neben der anfänglichen Testteilnahmemotivation auch der Verlauf der Testteilnahmemotivation innerhalb der Testsitzung modelliert werden. Auf diese Art können sowohl die intraindividuellen Veränderungen in Erwartung, Wert und Anstrengung während Tests als auch die interindividuellen Unterschiede in diesen intraindividuellen Veränderungen untersucht werden (Sayer & Cumsille, 2001). Zum Verständnis der interindividuellen Unterschiede in den intraindividuellen Veränderungen sollen zwei hypothetische Verläufe skizziert werden: a) Die Erfolgserwartung einer Person, die kommenden Aufgaben erfolgreich zu lösen, sinkt nach der Bearbeitung der ersten Testhälfte; b) Die Zuversicht einer anderen Personen steigt durch die Kenntnis der vorherigen Aufgaben an. Dies wäre der Extremfall interindividueller Unterschiede der intraindividuellen Veränderung (eine Abnahme vs. eine Zunahme der Erfolgserwartung). Ein weiterer Vorteil dieser Modelle liegt in der Testung, ob zu den unterschiedlichen Zeitpunkten dieselben Konstrukte erfasst wurden oder sich ihre psychometrischen Eigenschaften über die Zeit hinweg verändert haben. Die Testung der Messinvarianz über die Zeit stellt eine Voraussetzung der Anwendung dieser Art von Wachstumsmodell dar (Geiser, Keller & Lockhart, 2013).

Für die Beantwortung der ersten Fragestellung wird für die Erfolgserwartungen, Wichtigkeit des Tests und Anstrengungsbereitschaft jeweils ein Wachstumsmodell berechnet, um zu untersuchen, ob sich die drei Komponenten während der Testsitzung verändern. In einem zweiten Schritt werden die drei Wachstumsmodelle simultan berechnet, um die Zusammenhänge der Verläufe zu prüfen. Für die letzte Fragestellung werden die berechneten Wachstumsmodelle in ein latentes Regressionsmodell integriert. So kann der Zusammenhang zwischen Testleistung und der anfänglichen Testteilnahmemotivation sowie der Veränderung in der Testteilnahmemotivation (Erwartung, Wert und Anstrengung) überprüft werden. Darüber hinaus wird bei der Vorhersage der Testleistung für den sozio-demografischen Hintergrund der Schülerinnen und Schüler, deren domänenspezifische Motivation (analog zu Studie I) und Fähigkeit in Mathematik mithilfe der Mathematiknote kontrolliert.

Generell soll noch angemerkt werden, dass dies die erste Untersuchung von Testteilnahmemotivation in realen *large-scale low-stakes* Testsituationen ist, in der Schülerinnen und Schüler einen zweistündigen kognitiven *Paper-and-Pencil-Test* bearbeiten und in der alle drei Aspekte der Testteilnahmemotivation, nämlich Erwartung, Wert und Anstrengungsbereitschaft berücksichtigt werden.

# 4

# Studie I

## *The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences*

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-scale Assessments in Education, 2*(1), 1–17. doi:10.1186/s40536-014-0005-4

## Abstract

Low-stakes assessments do not have consequences for the test-takers. Currently, motivational research indicates that a lack of test-taking motivation can decrease students' performance in low-stakes assessments. However, little research has explored the domain-specific and situation-specific aspects of motivation simultaneously. Research examining differences in test-taking motivation among students in different types of schools is also limited. Our study therefore addressed the motivational determinants of test performance in low-stakes assessments, in general, as well as school-track-specific differences in particular. Drawing on national data from students who participated in a cross-national study of educational achievement, we conducted multiple regression analyses to predict the students' test performance and the effort they invested in that test. We conducted the analyses for the entire sample as well as for the students in that sample separated according to the school track they were attending. The results showed that, after we had controlled for self-concept in mathematics, test-taking motivation was significantly, but relatively weakly, associated with test performance: Students achieved higher test performance the more effort they invested and the less worry they experienced during the test. We also found school-track-specific differences for invested effort. Test attractiveness seems to be a more important inducement to invest effort for students in nonacademic-track schools than for students in academic-track schools. The weak relationship between test-taking motivation and performance supports the validity of the applied low-stakes test. However, it seems that invested effort and worry are indispensable constructs for performance in low-stakes tests. For students of nonacademic tracks especially, an attractive and enjoyable test seems a crucial aspect of motivating them to expend their best effort. Implications for constructing low-stakes tests are discussed.

*Keywords*: test-taking motivation; low-stakes assessments; effort; school-track-specific differences

# The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences

## 4.1 Introduction

Over the last two decades, an increasing number of education systems (hereafter countries) have found their participation in large-scale cross-national educational assessments a more and more relevant part of quality evaluation of their school systems. Examples of these studies are the Programme for International Student Assessment (PISA), conducted by the Organisation for Economic Co-operation and Development (OECD), and the Trends in International Mathematics and Science Study (TIMSS), conducted by the International Association for the Evaluation of Educational Achievement (IEA). Germany is one of the countries that regularly takes part in these comparative studies.

The results of these studies allow countries not only to assess how well their own students are performing, on average, but also to assess that performance against the average performance of students in the other participating countries. These rankings play an important role in government-led educational decision-making, which forms the basis for reforms. In order to draw valid conclusions about students' abilities during this process, test-takers need to be motivated to expend full effort throughout the entire testing session. However, such tests have no positive or negative consequences for the test-takers, no matter how successfully or unsuccessfully they perform.

These tests are often referred to as *low-stakes* tests. Accordingly, it is uncertain whether the students do actually expend full effort; it could be that the students' results do not depict their true level of ability due to low motivation. Therefore, the results of low-stakes assessments may not constitute a valid measure of students' abilities. In this case, a valid interpretation of the test results is threatened. Our aim in this study is to provide a closer look at the role of test-taking motivation on student performance in low-stakes assessments. Before describing our research questions in detail, we define test-taking motivation and provide an overview of previous research.

*Test-taking motivation*

Test-taking motivation is a specific type of achievement motivation that can be understood as an active process by which goal-oriented activity is initiated and maintained (Schunk, Pintrich, & Meece, 2008). It is assumed that students have domain-specific achievement motivation (e.g., motivation to engage in mathematics) and situation-specific achievement motivation (e.g., motivation to work hard in a specific school-based assessment). Domain-specific motivational constructs such as self-concept in mathematics cover a relatively stable personal trait, while situation-specific motivational constructs cover a state that can differ (e.g., depend on how the student feels "on the day"). Test-taking motivation is assigned to the latter motivational constructs, because taking a test is a specific situation for students. Baumert and Demmrich (2001) define this type of motivation as "the willingness to engage in working on test items and to invest effort and persistence in this undertaking" (p. 441).

In high-stakes tests, test-takers typically show high motivation to perform well because of the positive or negative consequences of their performance on that test (Barry & Finney, 2009). Research exploring test-taking motivation relative to low-stakes assessments presents a less clear picture. Most of these studies show a connection between test-taking motivation and performance on the one hand, and between test-taking motivation and test stakes on the other (Cole, Bergin, & Whittaker, 2008; Eklöf, 2007, 2008; Thelk, Sundre, Horst, & Finney, 2009; Wise & DeMars, 2005; Wolf & Smith, 1995). However, some studies have found no such relationships (Baumert & Demmrich, 2001; O'Neil, Abedi, Miyoshi, & Mastergeorge, 2005; O'Neil, Sugrue, & Baker, 1995). In the following subsections, we describe studies that have detected associations between test-taking motivation, performance, and test stakes, and those that have not. These studies include some of those just listed.

*Studies showing associations*

The investigation by Eklöf (2007, 2008) focused on the test-taking motivation of Swedish Grade 8 students in TIMSS 2003, deemed a low-stakes assessment, and examined both domain-specific and situation-specific aspects of motivation. In this study, the following motivational scales explained 31% of the variance in the students' average mathematics achievement scores: *mathematics self-concept* and *value of mathematics* as domain-specific factors of motivation as well as *test-taking motivation* as a situation-specific aspect

of motivation. Of these variables, mathematics self-concept was the most important predictor. However, after controlling for the domain-specific factors of motivation, Eklöf no longer found a significant relationship for the situation-specific aspect of motivation. Eklöf assumed that test-taking motivation had no effect because most of these Swedish Grade 8 students, having not previously experienced receiving grades or taken external tests, did not perceive the test as a low-stakes one.

Eklöf and Nyroos's (2013) analyses of data pertaining to performance of Grade 9 students on the Swedish national test of science achievement in 2009 supported the findings from the 2003 TIMSS data: a significant relationship between performance in science and (a) reported effort ($r$ = .25), (b) perceived importance of the test ($r$ = .20), and (c) test anxiety ($r$ = –.10). However, the authors could not consider the domain-specific aspects of motivation in their analyses because data on this matter were not collected during the assessment.

Cole et al. (2008) investigated the relationship of the following situation-specific aspects of motivation to the mathematics test performance of undergraduate students: interest, effort, and perceived usefulness and importance of the test. The results of the path analyses revealed that usefulness and importance of the test were strong predictors of effort (e.g., $R^2$ = .26 for mathematics), which in turn was an important predictor of test performance.

Lau, Swerdzewski, Jones, Anderson, and Markle (2009) tried to vary test-taking effort in a low-stakes assessment by changing the behavior of the test proctors (invigilators). The proctors were trained to point out the importance and usefulness of the test to the students and to encourage them to work hard. The proctors were also asked to create a productive working environment. The research team investigated the students' effort in testing sessions before (traditional sessions) and after implementation of the proctor-strategies (strategic sessions). Student effort was higher and less variable in the strategic sessions than in the traditional sessions (effect sizes between $d$ = 0.35 and $d$ = 0.57). The effect of increased effort on performance could not be analyzed because the tests before and after the implementation were slightly different in content, making performance on them noncomparable.

Other studies that have found a strong relationship between test-taking motivation and performance include those by Thelk et al. (2009) and Wise and DeMars (2005). The

latter two authors showed from their synthesis of 12 empirical studies that motivated students outperformed their unmotivated classmates by more than one-half of a standard deviation. However, Wise and DeMars cautioned that the relationship between test performance and test-taking motivation could have been distorted by academic ability as a mediator variable.

*Studies showing no associations*

One of the studies that found no relationship between test-taking motivation, performance, and test-stakes is that by O'Neil et al. (2005). They analyzed the effect of financial incentives on test-taking motivation and performance in mathematics, and divided their sample of test-takers into two groups. The Group 1 students were told they would receive a financial incentive of $10 per item correct. Also, in order to increase the credibility of the study, test-takers immediately received $20 if they got two simple items at the beginning of the test correct. Group 2 received no incentives for their participation. Group 1 reported significantly higher levels of test-taking effort and self-efficacy than Group 2 did. However, despite the high reward and the higher level of reported effort for the incentive group, there was no significant difference in performance between the treatment and the control group. The authors assumed that this outcome was due to the lack of correlation between effort and performance for the whole sample.

Similar results were found in a PISA 2000 pilot study in Germany (Baumert & Demmrich, 2001). The study examined whether increasing the test's stakes led to a higher level of test-taking motivation and a higher level of performance. Using an experimental design, the researchers manipulated the test conditions across four different groups of test-takers. The incentive for Group 1 was informational feedback, for Group 2 it was grades, and for Group 3 a financial reward. The fourth group was positioned as a reference group. Its members received the usual instructions accompanying PISA assessments and also had emphasized to them the social importance of tests in international comparative studies. In all groups, the invested effort was high, and the personal value of a successful test and the perceived usefulness of the test were the same. Furthermore, the authors found no treatment effects on test performance. While they considered many situation-specific aspects of motivation in their analyses, no domain-specific aspects of motivation were included.

*The importance of investigating school-track-specific differences for tracked school systems*

Before describing research on specific differences in motivation across types of school, which is one focus of our study, we consider it useful to explain Germany's tracked school system. After completing elementary school (grade 4 or grade 6, depending on the federal state), German students are assigned to different school tracks, primarily according to their scholastic performance. The academic track is the *Gymnasium*. The intermediate track has several school types, such as the *Realschule*, and the lower track is the *Hauptschule*. Of these school types, the *Gymnasium* (academic track) is the only one that exists in *all* German federal states.

One of the rationales for tracking in Germany is that school lessons can be better optimized according to student requirements if students are in homogeneous learning groups. For instance, because students in homogenous learning groups assumedly require similar learning time, groups with high achievers can cover more learning topics as well as topics with higher cognitive demands (Köller & Baumert, 2001, 2012). In short, the supposition is that students attain higher learning outcomes in homogeneous learning groups than in heterogeneous ones.

Significant differences in mean achievement occur across the schools in the three different tracks, while mean achievement in schools of the same track is generally similar (Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006). One investigation, for example, showed students in grade ten in academic-track schools outperforming students in intermediate-track schools and in lower-track schools in a mathematic test even after the researchers had controlled for math achievement in grade 7 at individual and school levels (Köller & Baumert, 2001). The differences between the schools in each track were only minor. Köller and Baumert suggested that one reason for the superior performance of the academic-track schools is because of their instruction culture, seen partly as a consequence of the teacher training (Köller & Baumert, 2001, 2012). Differentiation in student performance in any one school track or school will still occur, of course, commensurate with socioeconomic, psychosocial, motivational and cognitive variables. However, because achievement covaries with socioeconomic status to a very strong extent, social segregation is an undesirable ancillary effect of tracking. In essence, the different tracks "act" as developmental environments differentially influencing student performance (Baumert, Trautwein, & Artelt, 2003).

Reference to one of the studies already discussed in this paper—that by Baumert and Demmrich (2001)—is useful at this point. In addition to looking at the influence of incentives on test-taking motivation, the authors also compared the effort students in the lower-track *Hauptschule* and the academic-track *Gymnasium* put into their work on the particular test. The intended effort turned out to be the same for both school types, but the invested effort was lower for the *Hauptschule* students than for the *Gymnasium* students. The students in the academic-tracked schools reported a more positive emotional state and less task-irrelevant cognitions than the students in the lower-track schools. For the entire sample, self-reported effort and worry were the most powerful predictors of test performance. However, there was no investigation of the interplay between the motivational variables and their effects on performance for the different school tracks conducted.

Research on differences in domain-specific and trait-like motivational constructs across the school types has shown mixed results. Two investigations provide useful examples. Artelt, Demmrich, and Baumert (2001) found no differences across school tracks in students' mathematics self-concept or interest in the subject. The absence of self-concept differences suggests the "big fish little pond" effect may have been at play here (Marsh, 1987; Trautwein et al., 2006). According to this effect, students construct their self-concept by comparing themselves with their schoolmates; not by comparing themselves with all students of their age. Thus, students with a similar level of performance will report lower self-concepts if they are in a high-achieving environment (such as the academic track) than in a low-achieving environment. Consequently, despite students in academic-track schools knowing that their performance is higher than the performance of students in lower-track schools, they do not show a corresponding higher self-concept (Artelt et al., 2001).

In contrast, Baumert, Stanat, and Watermann (2006) found specific differences in students' self-efficacy beliefs across the school tracks. The authors used national data from the German extension sample of PISA 2000 to explore the influence of school structure on the emergence of differentiated learning environments. They also found evidence of the big fish little pond effect in that the self-efficacy beliefs of students with a similar level of achievement decreased as the track level of the school increased. Baumert and colleagues also conjectured that the larger proportion of class repeaters in the lower than higher tracks might lead to lower self-efficacy beliefs among the students in the lower-track schools.

Although this effect did not reach significance, the results nevertheless suggest that the concentration of underachievers in lower tracks can affect students' effort.

## 4.2 Study Objectives

The current state of research indicates that there is a relationship between test performance and test-taking motivation in low-stakes assessments. However, consideration of situation-specific and domain-specific aspects of motivation is lacking in most of the aforementioned studies. Moreover, the lack of research on school-track-specific differences in test-taking motivation and the mixed results of the cited studies on these differences points to the need for more investigation of test-taking motivation across school tracks. We therefore examined the relationship between different motivational aspects and students' performance in general and across school tracks in particular. Our initial research questions were the following:

1a) To what extent do domain-specific and situation-specific aspects of motivation predict students' performance in a low-stakes mathematics test?

1b) Are there school-track-specific differences in the relationship between performance in mathematics and domain-specific and situation-specific aspects of motivation?

In many studies, test-taking motivation is mainly operationalized through questions about students' invested effort, which covers the main element of the test-taking motivation definition. Accordingly, in a second step, we examined whether invested effort was influenced by other motivational aspects and again considered different school tracks in our research questions:

2a) To what extent do situation-specific aspects of motivation predict the invested effort of test-takers in a low-stakes test?

2b) Are there school-track-specific differences in the proportion of invested effort?

In summary, our research questions addressed two separate matters. The first focused on the relationship between performance and domain-specific as well as situation-specific aspects of motivation. The second focused on invested effort and its relationship with situation-specific aspects of motivation. Both sets of research questions also required us to consider differences in student performance across Germany's school tracks.

## 4.3 Method

*Participants*

We used the German extension sample of the PISA 2000 study (Deutsches PISA-Konsortium, 2003; Kunter et al., 2002; OECD & UNESCO Institute for Statistics, 2003) to investigate test-taking motivation and its relationship with students' performance in mathematics. The sample, nationally representative of German ninth-graders, consisted of 31,740 students. Half of the sample (50%) were female, and the average age of the students was 15.7 years (*SD* = 0.56). Thirty percent of the students were attending academic-track schools and 70% nonacademic-track schools. Eighty-eight percent of the students reported that German was their first language. The random sampling of schools was conducted by the IEA DPC (IEA Data Processing and Research Center), which is responsible for collecting PISA data in Germany.

*Procedure.* The PISA test took place in the spring of 2000. On the first day of testing, German students took the international standard assessment. On the following day, they took the national PISA extension assessment. The motivational questions used in our study were administered on this second day of testing. Students spent approximately three hours in total on the international and national administrations (two hours of performance tests and 30 minutes of student questionnaires accompanied by questions on cross-curricular competencies).

*Measurement instruments*

*Motivation subscale.* Although in PISA 2000, questions on test-taking motivation (situation-specific aspects of motivation) were administered before and after the test, we did not analyze test-taking motivation until after the test because the motivation scales varied slightly between the two measurements. For example, task-irrelevant cognition, one of the important predictor variables in the study by Baumert and Demmrich (2001), was not measured until the end of the test.

The post-test subscales assessed various aspects of test-taking motivation: emotional state, invested effort, test attractiveness, and usefulness of the test. The subscales were based on the items in the Online Motivation Questionnaire (Boekaerts & Otten, 1993). Items assessing task-irrelevant cognitions, namely worry and distraction, were also administered. These questions were derived from the Test Anxiety Inventory (Hodapp, Laux, & Spielberger, 1982). All self-reported items were measured on a four-point Likert

scale, with ratings ranging from 1 = strongly agree to 4 = strongly disagree. Negatively worded items within a positive scale were recoded. Table 4.1 provides examples of the items and also the subscales' internal consistencies.

Table 4.1

*Means, standard deviations, and internal consistencies of the test-taking motivation subscales*

| Aspect of motivation | Subscale | Item no. | Example item | *M* | *SD* | Internal consistency |
|---|---|---|---|---|---|---|
| Domain-specific | Self-concept in mathematics | 3 | I've always been good at math. | 2.48 | 0.94 | .89 |
| Situation-specific | Invested effort | 3 | How much effort you've given during the test? | 2.89 | 0.58 | .83 |
| | Test attractiveness | 3 | How much fun did you have during the test? | 2.56 | 0.62 | .82 |
| | Usefulness of the test | 1 | How useful do you find these kinds of tests? | 2.84 | 0.83 | - |
| | Emotional state | 4 | I'm in a good mood. | 2.87 | 0.67 | .81 |
| | Worry | 3 | I have doubted my abilities. | 2.06 | 0.69 | .73 |
| | Distraction | 2 | My thoughts wandered from the tasks. | 2.06 | 0.83 | .75 |

Although the subscales contained only a few items, the internal consistencies of the situation-specific subscales of motivation were all acceptable. The invested effort subscale assessed students' test-taking motivation defined according to Baumert and Demmrich's (2001) definition—willingness to engage on test items. We assigned the other subscales to the test-related facets (test attractiveness, usefulness of the test) and to the person-related facets (emotional state, worry, distraction) of test-taking motivation.

*Student background questionnaire.* Students completed this instrument with its self-report scales after they had taken the test. Among other constructs, this questionnaire, Marsh's (1990) Self Description Questionnaire, assessed students' self-concept in mathematics as a domain-specific aspect of motivation. Responses were measured on a four-point Likert scale, with ratings ranging from 1 = strongly disagree to 4 = strongly agree. Internal consistency was good (see Table 4.1). Studies by Brunner, Keller, Hornung, Reichert, and Martin (2009) and Chen, Yeh, Hwang, and Lin (2013) show that it is possible to distin-

guish both general and domain-specific dimensions of students' academic self-concept. We were mainly interested in our study in identifying any relationships among motivational constructs and test performance in mathematics, which is why we used only mathematical self-concept as the domain-specific component of academic self-concept.

*Achievement test.* The achievement test assessed reading, mathematical, and scientific literacy. In the present study, we drew on data from the national PISA test in mathematical literacy. The results were reported on an international scale with a mean of 500 and a standard deviation of 100. The PISA test is considered a low-stakes test because test-takers do not receive information about their performance and their results do not count towards their grades.

*Analyses*

In order to answer our research questions, we used Mplus 6 software (Muthén and Muthén, 1998-2010) to conduct multiple regression analyses with five plausible values (PVs). PVs are ability estimates, which we derived from an item response theory analysis (Yen & Fitzpatrick, 2006) conducted via ConQuest software (Wu, Adams, Wilson, & Haldane, 2007). Due to the structure of the student sample and the fact that students belonged to different classes, we used a clustering method to correct the standard errors. We also weighted the students for the population size. In order to gain a better interpretation of the results, we reported the unstandardized regression coefficient *b*, which reflects points on the international PISA achievement scale. In addition, because of the large sample size, we focused only on highly significant effects with a *p*-value below .001.

## 4.4 Results

Before presenting the findings pertaining to our research questions, we provide information about the students' test performance and their scores on the subscales of domain-specific and situation-specific motivational aspects. The weighted mean of the mathematics scores was 500.75 ($SD$ = 79.50). The standard deviation differed from the international metric because we computed the performance of the ninth-graders in the PISA German sample instead of all 15-year-olds in it.

As we anticipated, the academic-track students ($M_{at}$ = 573.83; $SD_{at}$ = 59.00) outperformed the nonacademic-track students ($M_{nt}$ = 470.30; $SD_{nt}$ = 65.95) on the mathematics test. As evident in Table 4.1, the students invested effort and concentrated on the items, enjoyed taking the test, and found the test useful. Accordingly, the students reported a

positive emotional state, little worry, and little distraction. The self-concept scores fell within a medium range, as did test attractiveness. The correlations between the several subscales ranged from $r = .00$ between worry and usefulness of the test to $r = .59$ between emotional state and test attractiveness. In summary, the pattern of students' ratings indicated that they were motivated to do well on the PISA 2000 test.

*Prediction of test performance with domain-specific and situation-specific aspects of motivation*

To answer our first research question (1a), the extent to which domain-specific and situation-specific aspects of motivation explained performance in the low-stakes PISA test, we examined the relationship between domain-specific and situation-specific aspects of motivation and test performance in mathematics. To accomplish this, we conducted a multiple regression analysis with mathematics performance as the criterion. The predictors were self-concept as the *domain-specific* aspect of motivation and the diverse test-taking motivation subscales as the *situation-specific* aspects of motivation. The procedure we used here followed the approach proposed by Eklöf (2008).

We added the predictors to the regression model in the following manner: first, self-concept as the domain-specific aspect of motivation; second, effort as the main element of test-taking motivation. We then added the test-related facets and the person-related facets, respectively. In a second step, we were interested in school-track-specific differences. Here, we conducted the regression analysis separately for students of two tracks: the academic-tracked schools (*Gymnasium*) and the nonacademic-tracked schools. Our decision to compare just two school tracks was because, as mentioned earlier, the *Gymnasium* is the only type of school that exists across all federal states.

Table 4.2 shows the results of the multiple regressions. In Model 1, self-concept in mathematics explained approximately 8% of the variance in mathematics scores. The regression coefficient ($b = 24.37$) was significant and indicated that an increase of 1 on the self-concept scale entailed an increase of approximately 24 score points on the PISA mathematics achievement scale. We then added invested effort as the first situation-specific aspect of motivation (Model 2). The variance explained increased slightly to 11%, and the effect of invested effort was significant. The third model included the test-related facets of test attractiveness and test usefulness. Usefulness had a significant but small coefficient, and the variance remained stable. The last model contained all aspects of

motivation. The overall explained variance was 15%, of which the domain-specific aspect of motivation, represented by students' self-concept in mathematics, contributed 8%: thus, the higher the students' self-concept in mathematics, the higher their performance in mathematics.

Table 4.2

*Multiple regression of mathematics performance on domain-specific and situation-specific aspects of motivation*

| | Model 1 | | Model 2 | | Model 3 | | Model 4 | |
|---|---|---|---|---|---|---|---|---|
| | *b* | *(SE)* | *b* | *(SE)* | *b* | *(SE)* | *b* | *(SE)* |
| Self-concept | 24.37* | (0.79) | 22.99* | (0.78) | 22.85* | (0.79) | 19.45* | (0.80) |
| Invested effort | | | 19.64* | (1.33) | 15.76* | (1.51) | 11.53* | (1.56) |
| Test attractiveness | | | | | 1.83 | (1.47) | -5.39 | (1.55) |
| Usefulness of the test | | | | | 4.78* | (0.97) | 5.93* | (0.96) |
| Emotional state | | | | | | | 1.04 | (1.26) |
| Worry | | | | | | | -19.65* | (1.12) |
| Distraction | | | | | | | -9.98* | (0.89) |
| $R^2$ *(SE)* | .08* | (0.01) | .11* | (0.01) | .11* | (0.01) | .15* | (0.01) |

*Note***.** *$p < .001$.

Overall, all subscales other than test attractiveness and emotional state significantly predicted test performance. The most important situation-specific aspects of motivation were (in order of size) worry, invested effort, distraction, and perceived usefulness of the test. These findings indicate that as the students' performance in mathematics improved, the (a) less worried they were, (b) more effort they invested, (c) less distracted they were, and (d) more useful they perceived the test to be. Thus, the situation-specific aspects of motivation—worry and distraction as well as invested effort—showed a relationship with performance, as did the domain-specific aspect of motivation, despite the small amount of variance that it explained.

Our second research question (1b) within this focus referred to the differences in the relationship between explained performance in mathematics and domain-specific and situation-specific aspects of motivation across school tracks. For a clearer presentation of the results, we chose only two models (shown in Table 4.3): the model with self-concept in

mathematics as the domain-specific aspects of motivation (Model 1), and the complete model with the domain- and situation-specific aspects of motivation (Model 4).

With the first model, we established differences between students in academic-tracked schools and students in nonacademic-tracked schools. Self-concept in mathematics explained 24% of the variance in the mathematics scores of the students in the first group of schools, but only 10% of the variance in the mathematics scores of the students in the second group.

Table 4.3

*Multiple regression of mathematics performance on domain-specific and situation-specific aspects of motivation, separated by school track*

| | Model 1 | | | | Model 4 | | | |
|---|---|---|---|---|---|---|---|---|
| | Nonacademic track | | Academic track | | Nonacademic track | | Academic track | |
| | *b* | *(SE)* | *b* | *(SE)* | *b* | *(SE)* | *b* | *(SE)* |
| Self-concept | 22.39* | (0.86) | 29.64* | (0.79) | 19.06* | (0.88) | 25.83* | (0.85) |
| Invested effort | | | | | 10.11* | (1.65) | 9.69* | (2.01) |
| Test attractiveness | | | | | -2.50 | (1.65) | 3.96 | (1.95) |
| Usefulness of the test | | | | | 1.10 | (1.06) | -2.15 | (1.19) |
| Emotional state | | | | | -2.32 | (1.31) | 3.50 | (1.68) |
| Worry | | | | | -16.64* | (1.16) | -12.78* | (1.31) |
| Distraction | | | | | -4.59* | (0.96) | -0.06 | (1.20) |
| $R^2$ *(SE)* | .10* | (0.01) | .24* | (0.01) | .15* | (0.01) | .29* | (0.01) |

*Note.* * $p < .001$.

In the complete model, the model to which we added the situation-specific aspects of motivation, the explained variance increased marginally, by 5%, for the students attending academic-track schools. Not only self-concept in mathematics but also worry and invested effort became relevant at this juncture, meaning that (a) the higher the self-concept of these students, (b) the less worried they were, and (c) the more they invested effort in the test, the better their performance on it.

The explained variance also increased marginally, again by 5%, in the complete model for the students in the nonacademic track. Here again, in addition to self-concept, the situation-specific aspects of motivation (i.e., worry and invested effort) were signifi-

cantly associated with performance. In contrast to the findings for the students in the academic-track, distraction also significantly predicted performance. Thus, for the students in the nonacademic-track schools, the higher their (a) self-concept and (b) invested effort, and the less their (c) worry and (d) distraction, the better they performed.

Our next step was to run a new regression for the complete model to determine if any of the interactions between motivational variables and school track were significant. Four of the seven interactions became statistically significant (ordered by size of the coefficients): test attractiveness ($b$ = 7.35), self-concept ($b$ = 5.95), emotional state ($b$ = 5.76), and distraction ($b$ = 4.17). We were surprised to find the interaction of emotional state and test attractiveness reaching significance given that the main effects of these two variables on test performance were not significant. We accordingly decided not to overemphasize these interactions given that these subscales did not seem to predict mathematics scores in either school track.

In summary, the interactions supported our results: self-concept and distraction demonstrated school-track-specific differences and were also significant predictors of test performance. The test-taking motivation scales explained the same amount of variance in performance in both school tracks.

*Prediction of invested effort with situation-specific aspects of motivation*

In order to answer the research question focusing on test-takers' invested effort in a low-stakes test by situation-specific aspects of motivation (2a), we used invested effort as criterion and the test-related (Model 1) and person-related facets of test-taking motivation (Model 2) as predictors. Because we were interested in the effects of situational aspects on effort, we did not include the domain-specific aspect of motivation. In a second step, we conducted a regression analysis, using the same approach as for the first set of research questions—that is, separately for the academic-track students and the nonacademic-track students.

Table 4.4 illustrates the results. In the first model, both of the test-related facets had a significant effect on invested effort whereby the coefficient of test attractiveness was bigger than the coefficient of usefulness of the test. Both explained 35% of the variance. In the complete model (i.e., containing the person-related facets), all subscales significantly predicted the invested effort and explained 40% of the variance in that effort. The test-related facets of test-taking motivation (test attractiveness, test usefulness) and distraction

had the strongest numerical values of the regression coefficients. The meaning that can be taken from this pattern is that the more (a) attractive and (b) useful the students perceived the test to be and (c) the less distracted they were, the higher their level of invested effort.

Table 4.4

*Multiple regression of invested effort on situation-specific aspects of motivation*

| | Model 1 | | Model 2 | |
|---|---|---|---|---|
| | *b* | *(SE)* | *b* | *(SE)* |
| Test attractiveness | 0.45* | (0.01) | 0.39* | (0.01) |
| Usefulness of the test | 0.13* | (0.01) | 0.10* | (0.01) |
| Emotional state | | | 0.04* | (0.01) |
| Worry | | | 0.06* | (0.01) |
| Distraction | | | -0.17* | (0.01) |
| $R^2$ *(SE)* | .35* | (0.01) | .40* | (0.01) |

*Note.* * $p < .001$.

The second question (2b) of this research focus referred to school-track-specific differences in the relationship between invested effort and situation-specific aspects of motivation. Table 4.5 contains the results. In Model 1, test attractiveness and test usefulness significantly predicted the invested effort for both school tracks. However, the coefficient of test attractiveness for nonacademic-track students was higher than for academic-track students. Correspondingly, the two test-related facets of test-taking motivation explained approximately 23% of the variance in the effort invested by the academic-track students, and 39% of the variance in effort invested by the nonacademic-track students.

Model 2, the model to which we added the person-related facets to the test-related facets, explained 31% of the variance in the effort the academic-track students invested in the mathematics assessment: test attractiveness, distraction, and usefulness of the test all showed significant coefficients. The pattern, then, was that (a) the more attractive and (b) useful the academic-track students perceived the test to be, and (c) the less distracted they were, the more effort they put into it. For nonacademic-track students, the complete model explained 43% of the variance in invested effort; all subscales significantly predicted that effort.

Table 4.5

*Multiple regression of invested effort on situation-specific aspects of motivation, separated by school track*

| | Model 1 | | | | Model 2 | | | |
|---|---|---|---|---|---|---|---|---|
| | Nonacademic track | | Academic track | | Nonacademic track | | Academic track | |
| | *b* | *(SE)* | *b* | *(SE)* | *b* | *(SE)* | *b* | *(SE)* |
| Test attractiveness | 0.49* | (0.01) | 0.35* | (0.01) | 0.43* | (0.01) | 0.28* | (0.02) |
| Usefulness of the test | 0.13* | (0.01) | 0.10* | (0.01) | 0.10* | (0.01) | 0.08* | (0.01) |
| Emotional state | | | | | 0.04* | (0.01) | 0.03 | (0.01) |
| Worry | | | | | 0.06* | (0.01) | 0.02 | (0.01) |
| Distraction | | | | | -0.16* | (0.01) | -0.19* | (0.01) |
| $R^2$ *(SE)* | .39* | (0.01) | .23* | (0.01) | .43* | (0.01) | .31* | (0.01) |

*Note.* * $p < .001$.

When we looked at the coefficients (those exceeding ± 10.0), we found that the pattern for the nonacademic-track students in Model 2 was similar to the pattern for the academic-track students. In order to assess whether the differences between the school tracks were statistically significant, we conducted a regression with interaction effects for the complete model. Again, as anticipated, the interaction between test attractiveness and school track showed a significant coefficient ($b$ = -0.15) as did the interaction between worry and school track ($b$ = -0.05). However, we do acknowledge that the latter coefficient is relatively small. In summary, the results relating to our second set of research questions suggests that the attractiveness of the test differed according to whether the students were from the academic-track schools or from the nonacademic-track schools.

## 4.5 Discussion

This study examined two sets of research questions focused on the relationship of various aspects of test-taking motivation, performance, and effort as well as school-track-specific differences within this relationship.

*Prediction of test performance with domain-specific and situation-specific aspects of motivation*

The first set of research questions examined the relationship between domain-specific and situation-specific aspects of motivation and test performance in mathematics. The results showed that nearly all situation-specific aspects of motivation predicted mathematics scores even after we had controlled for the domain-specific aspect of motivation. Self-concept as the domain-specific aspect of motivation explained slightly more variance than the situation-specific aspects of motivation. Along with self-concept, invested effort as well as worry and distraction as person-related facets of test-taking motivation had the greatest impact on the mathematics test scores (Research Question 1a).

These results do not support Eklöf's (2008) findings. In her study, test-taking motivation showed no significant effect on test performance when considered with domain-specific aspects of motivation. In order to explain these differences, we note that Eklöf (2008) examined a relatively small sample ($N$ = 343) of Swedish eighth-graders, whereas we used a nationally representative sample of German ninth-graders. As mentioned in the theoretical section of this paper, Eklöf assumed that students probably did not perceive the test as low-stakes because they had not yet experienced receiving grades or taking external tests. Hence, it is likely that test-taking motivation varies across countries due to cultural differences or different response behaviors. Thus, cross-country comparisons of test-taking motivation on low-stakes tests constitute an important area of further research.

Our results support the findings of Baumert and Demmrich (2001). In their study, as in ours, effort and worry were the most powerful predictors of test performance. However, these authors were able to explain nearly twice as much variance in performance on the basis of the two situation-specific aspects of motivation than we could with *all* of our domain- and situation-specific aspects of motivation. Unfortunately, they did not explicitly describe their analyses, which is why we were not able to compare these differences more concretely. Here, further research is necessary.

According to the school-track-specific differences, the results indicated that these differences are primarily due to the domain-specific aspect of motivation—students' self-concept of their mathematics ability. For academic-track students, self-concept had a stronger relationship with mathematics performance than it did for the students from the nonacademic track (Research Question 1b). These results correspond with the big fish little

pond effect (Marsh, 1987; Trautwein et al., 2006). Trautwein and colleagues concluded on the basis of their study that students construct their self-concept by comparing themselves with their schoolmates and not by comparing themselves with all students of their age. With respect to our study, this effect implies that even though students in the high-achieving environment (the *Gymnasium*) knew their achievement was higher on average than that of students in the lower-achieving environments, their self-concept was, on average, not higher than that of their lower-tracked peers.

Looked at another way, this pattern could mean that for the academic-track students, high self-concept actually corresponds with good performance (therefore the higher $R^2$), whereas for the students from the nonacademic track high self-concept does not necessarily lead to good performance (therefore the lower $R^2$). This hypothesis is supported by the correlation between self-concept and mathematics performance, which was higher for the academic-track students ($r$ = .48) than for their counterparts from the nonacademic track ($r$ = .32). When we used the Fisher's z test, we found this difference was highly significant ($z$ = -14.75, $p$ < .001).

The other school-track-specific difference we examined concerned the person-related facets of test-taking motivation. For nonacademic-track students, our findings suggest that it is more important that they do not doubt their abilities when taking tests and that they are focused on the tasks. With respect to the task-irrelevant cognitions, our findings correspond with the results of Baumert and Demmrich (2001), who found that academic-track students had a more positive emotional state and less task-irrelevant cognitions than students attending lower-track schools. Our results furthermore show that worry and distraction had a greater negative effect on performance for nonacademic-track students than for academic-track students. Thus, it is especially important that nonacademic-track students undergo testing in a distraction-free environment, with steps having been made to mitigate anxieties so that they are motivated to do their best. It may also be beneficial for further investigations to include questions assessing anxiety in their test-taking motivation scale (see, in this regard, Nie, Lau, & Liau, 2011; Putwain & Daniels, 2010).

*Prediction of invested effort with situation-specific aspects of motivation*

In regard to the second set of research questions, we found that the test- and person-related facets of test-taking motivation predicted invested effort, with test attractiveness emerging as the most powerful predictor. Distraction and usefulness of the test showed a smaller

relationship with invested effort (Research Question 2a), a finding that aligns with work by Cole et al. (2008). They found that perceived usefulness and importance of the test were strong predictors of effort. For the nonacademic-track students, test attractiveness was more relevant than for the academic-track students (Research Question 2b); a positive image of low-stakes assessments and a calm working atmosphere appear to have been essential aspects of a favorable test environment for this first group of students. This finding can be regarded as "good news" because it suggests that test-related facets of test-taking motivation can be positively influenced by making low-stakes tests interesting and appealing. Even if the test has no consequences for the test-takers, it is nonetheless important that they find it an enjoyable experience.

Our study furthermore found that performance in low-stakes tests was slightly influenced by different motivational aspects of test-taking motivation. These results imply that students are likely to achieve higher test performance the more effort they invest and the less worry they experience during the testing session. These small effects support the general validity of this low-stakes assessment in Germany. Thus, educational policy decision-making processes based on the results of low-stakes assessments can be supported for this sample. However, we do not know whether the small effects depend on the country in which it is administered, or on the particular test. It is thus important to take into account motivational measures, such as students' invested effort, when endeavoring to draw valid conclusions about students' performance. The school-track-specific differences in self-concept in mathematics and in invested effort that we found imply that for students attending nonacademic-tracked schools especially, an attractive and enjoyable test is crucial to motivate them to do their best. This consideration should be kept in mind by researchers when constructing low-stakes tests items.

*Limitations and Conclusion*

A limitation of the present study is the number of items per subscale. For example, the test usefulness subscale had just one item, while the distraction subscale contained only two. The internal consistencies of these two subscales could be improved by adding further items to them. Due to the restricted testing time and the large number of questions in the student questionnaire, more items could not be implemented. However, good and substantial reliabilities confirmed the homogeneity of these scales.

Another limitation concerns the fact that the students completed the full motivational questionnaire after they took the test. Thus, it is possible that their responses to the self-report questionnaire were confounded by their perceived test performance. According to attribution theory, it seems likely that students reported lower invested effort to justify their lower perceived test performance (Weiner, 1986). Whether or not the reported level of test-taking motivation corresponded with the actual test-taking motivation during the test is therefore uncertain. We intend to undertake further research to explore reported test-taking motivation before a test and its relationship with test performance. We also intend to compare reported test-taking motivation before a test with the test-taking motivation after it using the same motivational subscales.

In general, further investigations similar to the experimental study of Baumert and Demmrich (2001) are necessary. They found no effect of raising the stakes on effort and performance. However, their study was conducted before the first PISA survey, which was administered in 2000. Over the intervening years, the frequency of international and national tests in German schools has greatly increased; today, students take more external tests than they did at the beginning of this century. Thus, after more than a decade of intense testing, it seems likely that test-taking motivation in low-stakes assessments has developed an influence on effort and performance. Just how motivated students remain throughout the testing session is another area of particular interest. An analysis of this kind would rely on more than just two measurements (i.e., before and after the test). Once such data are to hand, the course of students' test-taking motivation during testing sessions can be more robustly examined.

## References

Artelt, C., Demmrich, A., & Baumert, J. (2001). Selbstreguliertes Lernen: Motivation und Strategien in den Ländern der Bundesrepublik Deutschland [Self-regulated learning: Motivation and strategies in the German federal states]. In Deutsches PISA-Konsortium (Ed.), *PISA 2000: Basiskompetenzen von Schülerinnen und Schülern im internationalen Vergleich* (pp. 271–298). Opladen: Leske + Budrich.

Barry, C. L., & Finney, S. J. (2009). *Exploring change in test-taking motivation*. Paper presented at the annual meeting of the Northeastern Educational Research Association, Rocky Hill, CT.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441–462.

Baumert, J., Stanat, P., & Watermann, R. (2006). Schulstruktur und die Entstehung differenzieller Lern- und Entwicklungsmilieus [School structure and the emergence of differential learning and developing milieus]. In .J Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit Vertiefende Analysen im Rahmen von PISA 2000* (pp. 95–188). Wiesbaden: VS Verlag für Sozialwissenschaften/GWV Fachverlage GmbH, Wiesbaden.

Baumert, J., Trautwein, U., & Artelt, C. (2003). Schulumwelten: institutionelle Bedingungen des Lehrens und Lernens [School environments: Institutional conditions of teaching and learning]. In Deutsches PISA-Konsortium (Ed.), *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland* (pp. 261–331). Opladen: Leske + Budrich.

Boekaerts, M., & Otten, R. (1993). Handlungskontrolle und Lernanstrengung im Schulunterricht [Action control and learning-related effort in the classroom]. *Zeitschrift für Pädagogische Psychologie, 7*(2/3), 109–116.

Brunner, M., Keller, U., Hornung, C., Reichert, M., & Martin, R. (2009). The cross-cultural generalizability of a new structural model of academic self-concepts. *Learning and Individual Differences, 19*(4), 387–403. doi:10.1016/j.lindif.2008.11.008

Chen, S.-K., Yeh, Y.-C., Hwang, F.-M., & Lin, S. S. J. (2013). The relationship between academic self-concept and achievement: A multicohort–multioccasion study. *Learning and Individual Differences, 23*, 172–178. doi:10.1016/j.lindif.2012.07.021

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*(4), 609–624.

Deutsches PISA-Konsortium (2003). *PISA 2000 – Ein differenzierter Blick auf die Länder der Bundesrepublik Deutschland [PISA 2000 – A differentiated view of the German federal states]*. Opladen: Leske + Budrich.

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*(3), 311–326.

Eklöf, H. (2008). Test-taking motivation on low-stakes tests: A Swedish TIMSS 2003 example. In *Issues and methodologies in large-scale assessments: IERI monograph series* (Vol. 1, pp. 9–21). Hamburg: IEA-ETS Research Institute.

Eklöf, H., & Nyroos, M. (2013). Pupil perceptions of national tests in science: Perceived importance, invested effort, and test anxiety. *European Journal of Psychology of Education, 28*(2), 497–510. doi:10.1007/s10212-012-0125-6

Hodapp, V., Laux, L., & Spielberger, C. D. (1982). Theorie und Messung der emotionalen und kognitiven Komponente der Prüfungsangst [Theory and measurement of emotional and cognitive component of test anxiety]. *Zeitschrift für Pädagogische Psychologie, 3*(3), 169–184.

Köller, O., & Baumert, J. (2001). Leistungsgruppierungen in der Sekundarstufe I [Performance grouping in secondary education]. *Zeitschrift für Pädagogische Psychologie, 15*(2), 99–110. doi:10.1024//1010-0652.15.2.99

Köller, O., & Baumert, J. (2012). Schulische Leistung und ihre Messung [School achievement and its measurement]. In W Schneider & U Lindenberger (Eds.), *Entwicklungspsychologie* (Vol. 7, pp. 645–661). Weinheim: Beltz/PVU.

Kunter, M., Schümer, G., Artelt, C., Baumert, J., Klieme, E., Neubrand, M., Prenzel, M., Schiefele, U., Schneider, W., Stanat, P., Tillmann, K.-J., & Weiß, M. (2002). *PISA 2000: Dokumentation der Erhebungsinstrumente (Bd. 72) [PISA 2000: Documentation of the survey instruments]*. Berlin: Max-Planck-Institut für Bildungsforschung.

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education, 58*(3), 196–217. doi:10.1353/jge.0.0045

Marsh, H. W. (1987). The big-fish-little-pond effect on academic self-concept. *Journal of Educational Psychology, 79*(3), 280–295. doi:10.1037/0022-0663.79.3.280

Marsh, H. W. (1990). *Self Description Questionnaire (SDQ) II: A theoretical and empirical basis for the measurement of multiple dimensions of adolescent self-concept: An interim test manual and a research monograph*. San Antonio, TX: The Psychological Corporation.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide. Sixth edition*. Los Angeles, CA: Muthén & Muthén.

Nie, Y., Lau, S., & Liau, A. K. (2011). Role of academic self-efficacy in moderating the relation between task importance and test anxiety. *Learning and Individual Differences, 21*(6), 736–741. doi:10.1016/j.lindif.2011.09.005

OECD & UNESCO Institute for Statistics (2003). *Literacy skills for the world of tomorrow: Further results from PISA 2000*. Paris: OECD Publishing.

O'Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment, 10*(3), 185–208. doi:10.1207/s15326977ea1003_3

O'Neil, H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment, 3*(2), 135–157.

Putwain, D. W., & Daniels, R. A. (2010). Is the relationship between competence beliefs and test anxiety influenced by goal orientation? *Learning and Individual Differences, 20*(1), 8–13. doi:10.1016/j.lindif.2009.10.006

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). *Motivation in education: Theory, research, and applications (3rd ed.)*. Upper Saddle River, NJ: Pearson Education.

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the student opinion scale to make valid inferences about student performance. *The Journal of General Education, 58*(3), 129–151. doi:10.1353/jge.0.0047

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology, 98*(4), 788–806. doi:10.1037/0022-0663.98.4.788

Weiner, B. (1986). *An attributional theory of motivation and emotion*. New York: Springer.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. doi:10.1207/s15326977ea1001_1

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*(3), 227–242. doi:10.1207/s15324818ame0803_3

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0: Generalised item response modelling software*. Camberwell, VIC: ACER Press.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement*, (4th ed., pp. 111–153). Westport, CT: Praeger Publishers.

# 5

# Studie II

## *Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation*

Penk, C., & Schipolowski, S. (2015). Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation. *Manuscript submitted for Publication*.

(Stand: Februar 2015)

## Abstract

We investigated test-taking motivation in a large-scale assessment context by applying expectancy-value theory as the framework most commonly used to conceptualize test-taking motivation. Specifically, our aim was to explore the complex relationship between expectancy, value, test-taking effort, and test performance using data from a large-scale educational assessment study of German ninth-graders. First, we established a measurement model of test-taking motivation including all aspects of this multidimensional construct. Second, we investigated the predictive power of different facets of test-taking motivation for test-taking effort and test performance. Factor analyses indicated that expectancy, value, and test-taking effort constitute distinguishable components of test-taking motivation. Subsequent latent regression analyses showed that the value component was a strong predictor of test-taking effort and that expectancy, value, and effort taken together explained over a quarter of the variance in mathematics scores. Expectancy and test-taking effort had the most pronounced effects on test performance. We conclude that a comprehensive model of test-taking motivation should include all three components, that is, expectancy, value, and test-taking effort. Implications for future research are discussed.

*Keywords:* expectancy-value theory, test-taking motivation, test-taking effort, low-stakes tests, large-scale assessment, mathematics achievement

# Is it all about value? Bringing back the expectancy component to the assessment of test-taking motivation

## 5.1 Introduction

Expectancy-value theory is one of the most important conceptions of achievement motivation. The theory assumes that expectancies and values directly affect achievement behavior, such as test performance, as well as effort, choice, and persistence (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000). The expectancy-value model is also the framework most commonly used to conceptualize test-taking motivation, which is a particular type of achievement motivation. However, although the expectancy component is assumed to be a better predictor of test performance than the value component (Schunk, Pintrich, & Meece, 2008), most research on test-taking motivation considers only the value component (Butler & Adams, 2007; Cole, Bergin, & Whittaker, 2008; Eklöf & Nyroos, 2013; Eklöf, Pavešič, & Grønmo, 2014; Swerdzewski, Harmes, & Finney, 2011; Wolf & Smith, 1995). Furthermore, the few investigations including both expectancy for success and value of the test often fail to consider test-taking effort, which is considered the main element in the definition of test-taking motivation (Asseburg, 2011; Freund & Holling, 2011; Freund, Kuhn, & Holling, 2011).

The aim of the present study was to investigate *both* components of expectancy-value theory *and* test-taking effort in a large-scale assessment context. Specifically, we examined the complex relationship between all aspects of the test-taking motivation construct (including expectancy, value, and test-taking effort) and test performance using data from a large-scale educational assessment study of German ninth-graders. We pursued three objectives: a) establishing a measurement model including all aspects of test-taking motivation, b) predicting test-taking effort, and c) predicting test performance. Before defining our research questions in more detail, we describe the expectancy-value theory and provide an overview of the current state of research on test-taking motivation.

*Expectancy-value theory as a framework for test-taking motivation*

The expectancy-value theory provides a theoretical framework for test-taking motivation (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000). Based on Atkinson's (1957, 1964) theory of achievement motivation, the main components of the expectancy-value theory – expectancy for success and the perceived value of a task – are assumed to affect achievement behavior, that is performance as well as effort, persistence, and the decision for (or

against) taking on a task. The expectancies refer to the students' beliefs of how well they will perform, and therefore include the individual's perception of his or her own competence at a given task. The value component consists of four distinct aspects: attainment value (importance), intrinsic value (enjoyment), utility value (usefulness of the task), and cost (emotional cost or effort). The value aspects are assumed to explain performance-related decisions based on students' beliefs about how they might benefit from a task. By comparison, the expectancy component has been shown to be a stronger predictor of performance while the value component is more closely associated with persistence or choice than the expectancy component (Eccles & Wigfield, 2002; Schunk et al., 2008; Wigfield & Eccles, 2000; Wigfield, 1994). The classification of effort in the expectancy-value model of Eccles and colleagues is ambiguous. On the one hand, effort is defined as an outcome of both expectancy and value. On the other hand, it is associated with the cost aspect of the value component because cost is construed in terms of how much effort is needed to succeed (Eccles & Wigfield, 2002).

Test-taking motivation can be regarded as a special state equivalent of the trait-like achievement motivation and is defined as "the willingness to engage in working on test items and to invest effort and persistence in this undertaking" (Baumert & Demmrich, 2001, p. 441). The expectancy-value theory of achievement motivation provides an appropriate framework in the context of test-taking motivation because it is one way to explain the relationship between motivation and test performance in low-stakes assessments. Low-stakes assessments represent a special situation for the test-takers as their performance in such tests has no personal consequences for them. Therefore, one could ask why the students should perceive a low-stakes test as important or useful and why they should put forth effort to complete the test. Given the inherent low value of low-stakes tests the expectancy-value theory explicates why students report low levels of test-taking effort, and in turn, may have lower test performance than in high-stakes testing situations. Therefore, test-taking motivation is often assessed with questions about effort and importance of the test. On a political level, however, such tests are high-stakes because they provide crucial information on the relative strengths and weaknesses of the educational system in a country (Stanat & Lüdtke, 2013).

In the context of expectancy-value theory, our conceptualization of test-taking motivation in the present paper is composed of three constructs: expectancy, value, and effort. Thus, test-taking effort is included in the expectancy-value model as a third main compo-

nent besides expectancy and value. Effort is defined as "a student's engagement and expenditure of energy toward the goal of attaining the highest possible score on the test" (Wise & DeMars, 2005, p. 2) and constitutes an important construct because the students are asked to expend effort even though the test results have no immediate consequences for them. In line with Wigfield and Eccles (2000), we assume that effort is an outcome of expectancy and values and is related to test performance for the remainder of this article.

*Previous research on test-taking motivation*

In this section we describe the current state of research with regard to the three components of test-taking motivation: expectancy, value, and effort. All studies described in the following, except Eklöf and Nyroos (2013), were undertaken in low-stakes contexts and explored whether test-taking motivation predicted test performance. First, we consider studies that assessed effort and value followed by studies that measured expectancy and value.

Most research on test-taking motivation considers test-taking effort and the value component (Thelk, Sundre, Horst, & Finney, 2009) but neglects expectancy for success, assuming that test-takers cannot conceive of 'success' in tests without any feedback or consequences (Cole et al., 2008). For instance, Baumert and Demmrich (2001) explored test-taking motivation in a pilot study of the first *Programme for International Student Assessment* (PISA) survey. In this study, self-reported effort and worry were the most powerful predictors of test performance and taken together explained 28% of the variance in mathematics achievement. The effects of the other motivational variables on test performance, that is, personal value of a successful test and the perceived usefulness of the test, were mediated by effort and worry. Eklöf and Nyroos (2013) investigated test-taking motivation of students in the first *Swedish National Test* of 2009 assessing biology, chemistry, and physics. The stakes of this test were "semi-low" because teachers could use the test results as part of their student evaluations but this was not obligatory. The data showed a significant relationship between performance in science and test-taking effort ($r = .25$) as well as between test performance and different aspects of the value component, specifically, perceived importance of the test ($r = .20$) and test anxiety ($r = –.10$). Similar results were obtained for the Swedish TIMSS 2003 data, a typical low-stakes test (Eklöf, 2007). However, in a multiple regression of test performance on test-taking motivation, mathematical self-concept, and value of mathematics using TIMSS 2003 data, test-taking motivation did not significantly predict performance. A total of 31% of the variance in

mathematics scores could be explained in the regression analysis, but this was largely due to individual differences in domain-specific perception of competence. Cole, Bergin, and Whittaker (2008) explicitly focused on the value component in a study with undergraduate students in the following subjects: English, Math, Science, and Social studies. The study assessed three aspects of the value component (interest, usefulness, and importance of the test) and effort. Path analyses revealed a similar pattern of results for all subjects: Usefulness and importance of the test were strong predictors of effort (e.g., $R^2 = .26$ for mathematics) which, in turn, was an important predictor of test performance. The authors found that effort fully mediated the effect of importance and usefulness of the test on performance in mathematics. The aspects of the value component and effort taken together had a similarly strong effect on performance as the ACT exam score, a standardized test for college admission (Cole et al., 2008). Recent research supports the mediating role of test-taking effort: In a study by Zilberberg, Finney, Marsh, and Anderson (2014), perceived importance of the test was a significant predictor of test-taking effort which, in turn, significantly predicted test performance. Zilberberg and colleagues controlled for gender and quantitative ability which may explain the rather small magnitude of the indirect effect ($\beta = .09$). However, they did not consider the remaining three value aspects (i.e., intrinsic value, utility value, and cost) or the expectancy component.

Other studies considered both the expectancy and the value component but ignored test-taking effort. For instance, Asseburg (2011) explored the relationship between expectancy for success and perceived value in a low-stakes study with German ninth-graders. In this study, a large number of constructs was measured beyond expectancy and value (i.e., importance of the test) such as ability beliefs (i.e., self-concept and self-efficacy), hope of success, and perceived test performance. The expectancy component explained about 10% of the variance in actual mathematics performance whereas the value component was not a significant predictor of test performance. As a consequence, Asseburg (2011) recommended that an ‘expectancy-model’ of test-taking motivation is adequate for low-stakes test situations. However, test-taking effort was not measured. Freund, Kuhn, and Holling (2011) assessed motivation of university and secondary school students before taking an abstract reasoning test. The authors used the Questionnaire on Current Motivation (QCM) which captures the expectancy component with the scale ‘probability of success’ and the value component with the scales ‘challenge’, ‘interest’, and ‘anxiety’ (Rheinberg, Vollmeyer, & Burns, 2001). Interest was the only significant

predictor of test performance; the effect of probability of success failed to reach statistical significance ($p$ = .07). Taken together, all motivation-related scales explained 14% of the variance in the test scores (Freund et al., 2011). These results were confirmed by Freund and Holling (2011) who examined the relationship between expectancy, value, and test performance. They also administered the QCM before the test and found that interest and perceived probability of success significantly predicted scores on a figural reasoning test after accounting for individual differences in general mental ability. Altogether, test-taking motivation and general mental ability explained 32% of the variance in the test scores. In addition, they conducted a retest seven to 14 days after the first measurement. Current motivation was assessed again before the students took the retest. All QCM scales were significant predictors of test performance and together with general mental ability explained over half of the variance in the test scores. Remarkably, probability of success was the strongest predictor of performance, followed by general mental ability. The authors concluded that if the test-takers gain experience with the ability test, the importance of test-taking motivation in predicting test performance increases (Freund & Holling, 2011).

The last two studies mentioned above assessed test-taking motivation before administering the ability test. All other studies assessed test-taking motivation after the participants had completed the test, therefore measuring the invested test-taking motivation. Thus, the studies discussed here differ in terms of both the test-taking motivation constructs that were assessed and the time point of measurement (i.e., before or after the ability test). It remains unclear whether the disparate results can be attributed to the different test-taking motivation constructs assessed in the studies (i.e., expectancy, value, or effort) or due to the different time points of the test-taking motivation measurement.

## 5.2 Study objectives

In sum, although the studies discussed above were based on typical low-stakes tests, they obtained diverging results. The first set of studies analyzed the value component and test-taking effort but did not consider the expectancy component of test-taking motivation. In contrast, the last set of studies explored both components of the expectancy-value model but failed to assess test-taking effort. Given the potential of expectancy-value theory for predicting test-taking motivation and test performance, however, it is important to explore *all* aspects of the test-taking motivation construct. This is the primary aim of the present study in which we investigated expectancy, value, *and* test-taking effort in terms of their interrelations and their relation to test performance in a large-scale low-stakes assessment.

Moreover, it is unclear if the time point of test-taking motivation measurement has an influence on the effect of test-taking motivation on performance. For this reason, we explored both test-taking motivation measured before the test and test-taking motivation measured after the test.

In a first step, we established a measurement model for all test-taking motivation constructs and analyzed their interrelations. In the next step, we conducted latent regression analyses to a) predict test-taking effort with expectancy for success and perceived value of the test, and to b) predict actual test performance with all components of test-taking motivation (expectancy for success, perceived value of the test, and test-taking effort).

For the prediction of test-taking effort with perceived value of the test and expectancy for success (Model 1, see Figure 5.1), our research questions were as follows:

1a) Does the value component explain students' level of test-taking effort?

1b) Does the expectancy component explain students' level of test-taking effort beyond the value component?

We included value of the test in the first model because it is the only construct considered in all previous studies. On the basis of the theoretical assumptions of the expectancy-value model, we assumed that both expectancy and value contribute significantly and similarly to the prediction of effort.

In the prediction of test performance with perceived value of the test, expectancy for success, and test-taking effort, we investigated a possible mediator effect of test-taking effort (see Model 2 in Figure 5.1), corresponding with Cole et al. (2008). Here, our research questions were as follows:

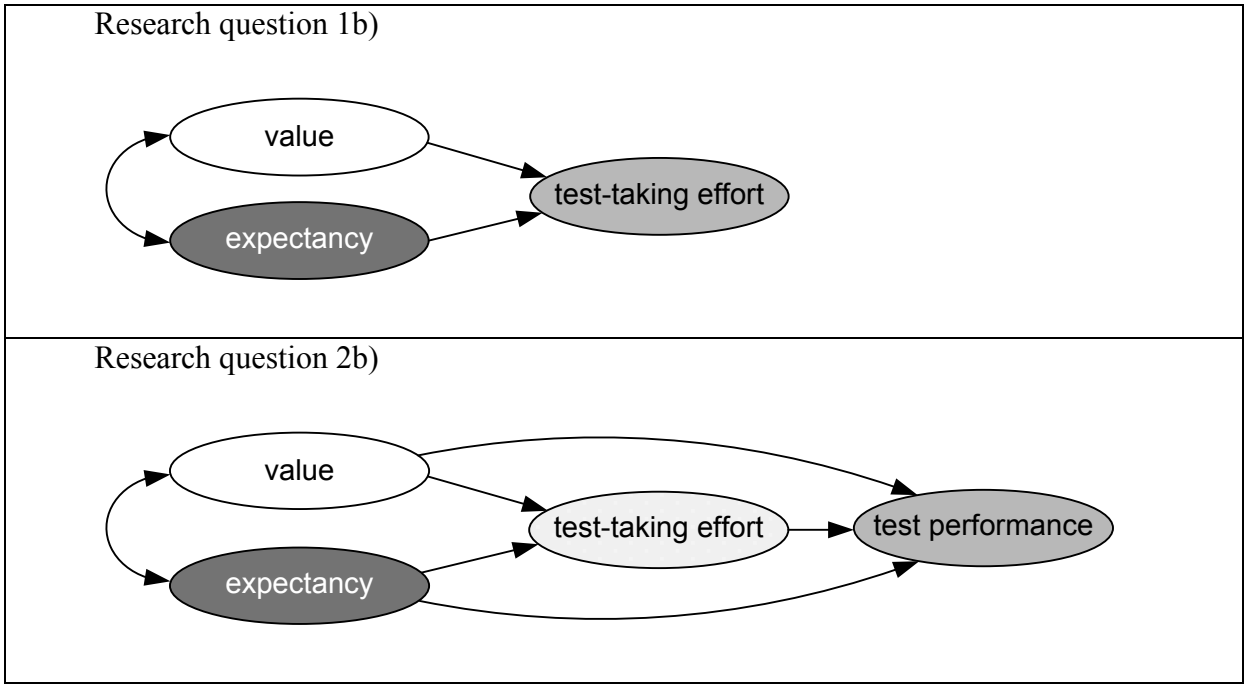
2a) Does test-taking effort explain student test performance?

2b) Do the expectancy and value components explain student test scores beyond test-taking effort (i.e., are expectancy and value incrementally valid predictors of test performance)? Are the effects of the value and expectancy components on test performance mediated by test-taking effort?

On the basis of previous research (Eklöf et al., 2014, Thelk et al., 2009; Wise & DeMars, 2005) we assumed that test-taking effort is significantly related to test performance. Second, we hypothesized that the effects of both the expectancy component and the value

component on test performance are partially mediated by test-taking effort. Third, we assumed that the expectancy component predicts test performance beyond test-taking effort (i.e., that there is a significant direct effect of the expectancy component on test performance). The last hypothesis bears on the work of Eccles and colleagues who found a stronger relationship between expectancy for success and performance than between perceived value and performance (Eccles & Wigfield, 2002; Wigfield & Eccles, 2000).

Additionally, we were interested in whether test-taking motivation before the test differs from test-taking motivation after the test. Therefore, we explored the research questions mentioned above for test-taking motivation measured before and after the test. Although exploratory, this research question is an important contribution to the discussion whether test-taking motivation before and after the test have different effects on test performance.




*Note.* Manifest indicators and disturbance terms are omitted for simplicity.

*Figure 5.1*. Overview of theoretical models of the research questions.

## 5.3 Method

*Sample*

We measured test-taking motivation in a large-scale educational study of mathematical and scientific literacy in Germany (Pant et al., 2013). The sample was nationally representative for German ninth-graders and consisted of 44,584 students. We excluded students with

special needs ($n$ = 1,380) because they received a shorter and easier achievement test. In addition, we excluded students who intentionally disregarded the instructions of the test-taking motivation questionnaire (i.e., handing in a completely blank questionnaire or ticking the same response alternative on all items; $n$ = 906). A total of 42,298 students constituted the sample used for all further analyses. Almost half of the sample was female (49.8%) and the mean age was 15.6 years ($SD$ = 0.61).

*Procedure and Instruments*

*Procedure*. The assessment was carried out in spring 2012. Test-taking motivation was assessed before and after the achievement test. After presenting general instructions for the test, the test-taking motivation items were presented to all students (version: before the test). Following the test-taking motivation questionnaire, the students completed the achievement test which took two hours plus a break of 15 minutes after the first hour. After the test, a student questionnaire was presented that assessed demographic information and attitudes towards school. For approximately half of the sample, this questionnaire also contained items on test-taking motivation (version: after the test).

*Test-taking motivation*. To measure both components of the expectancy-value theory, we used the Questionnaire on Current Motivation (QCM, Freund et al., 2011; Rheinberg, Vollmeyer, & Burns, 2001). The questionnaire contained four scales with three items per scale: challenge (“If I can do this test, I will feel proud of myself”), interest (“For tests like this I do not need a reward, they are lots of fun anyhow”), probability of success (“I think I am up to the difficulty of this test”), and anxiety (“I feel under pressure to do this test well”), with all items using a Likert scale ranging from *1 = strongly disagree* to *4 = strongly agree*. The expectancy component was represented by perceived probability of success; the value component was represented by the remaining three scales. In more detail, the concept of challenge denotes the extent to which test-takers perceive the situation as an achievement situation. Note that the challenge scale is henceforth referred to as the importance scale to prevent misinterpretation. Probability of success refers to the confidence of the test-takers to do well on the test. Interest refers to the degree to which the test-takers value the test content as well as the voluntariness of test completion. Anxiety describes the fear of failure and the state of feeling pressured in the achievement situation and, therefore, represents the cost aspect of the value component (Freund et al., 2011; Rheinberg et al., 2001).

All scales related to the current test situation and thus aimed to measure states. Originally, the QCM was developed to assess current motivation before taking a cognitive test. To employ the scales for an assessment of test-taking motivation after the test, we adapted the items accordingly (i.e., for a retrospective test-taking motivation measurement).

Additionally, we assessed test-taking effort at both time points with three items of the test-taking motivation scale by Eklöf (2010). This scale was originally applied after a test ("I felt motivated to do my best on this test"); we adapted the items for the first time point to measure test-taking effort before the test ("I feel motivated to do my best on this test").

*Achivement test*. The achievement test (*Ländervergleich 2012;* Pant et al., 2013) consisted of 374 mathematics items (39% multiple choice, 11% constructed response, and 50% open-ended) and 386 science items (59% multiple choice, 19% constructed response, and 22% open-ended). As is typical of educational large-scale assessments, a balanced incomplete block design was used (i.e., every student was administered only a selection of the items; Frey, Hartig, & Rupp, 2009). Specifically, for each of the domains, the items were distributed across 31 booklets. The test was based on national educational standards and was aimed at evaluating and comparing math and science competencies of students in the German federal states; therefore, it was a typical low-stakes assessment for the test-takers who were not graded and did not receive individual feedback about their performance.

*Analyses*

Analyses were conducted using confirmatory factor analysis (CFA) and latent variable regression analysis in Mplus (version 7; Muthén & Muthén, 1998-2012). Weighted Likelihood Estimates (WLEs; Warm, 1989) based on unidimensional scaling with the 1-parameter logistic (Rasch) model were used as manifest indicators of mathematics achievement. The hierarchical structure of the data was taken into account in the computation of standard errors and model fit. Also, case weights were used in all analyses so that the results were generalizable to the population of ninth-graders in German schools. Due to the large sample size, we specified a *p*-value below .001 as the cut-off for statistical significance.

First, a measurement model of the five test-taking motivation scales was developed to test if importance, interest, anxiety, probability of success, and test-taking effort

constitute distinguishable components of test-taking motivation. Second, we explored the relationships between the five scales and test performance by estimating a series of four latent regression models. Table 5.1 gives an overview of the constructs included in each regression model[1]. Note that due to the different sets of variables included in the models, Model 1a is not nested within Model 1b; the same applies to models 2a and 2b. We analyzed the various motivational constructs before and after the test to compare students' reported test-taking motivation before and after the test.

In all analyses, we applied a robust maximum likelihood (MLR) estimator and considered the following indices to evaluate model fit: 1) MLR $\chi^2$-statistic, corresponding degrees of freedom, and probability value (note that the $\chi^2$ value using the MLR estimator is asymptotically equivalent to the Yuan-Bentler T2* test statistic; Muthén & Muthén, 1998-2012), 2) comparative fit index (CFI), 3) Tucker–Lewis index (TLI), 4) root mean square error of approximation (RMSEA), and 5) standardized root mean square residual (SRMR). According to Hu and Bentler (1999), the following values indicate adequate model fit: CFI > .95, TLI > .95, RMSEA < .06, and SRMR < .08. Full Information Maximum Likelihood (FIML) estimation was used to handle missing data (Enders, 2010).

Table 5.1

*Specification of latent variable regression models*

| | | | Model | | | |
|---|---|---|---|---|---|---|
| | Component | Measures | 1a | 1b | 2a | 2b |
| Test-taking motivation | Value (v) | Importance | p | p | - | p |
| | | Interest | p | p | - | p |
| | | Anxiety | p | p | - | p |
| | Expectancy for success (e) | Probability of success | - | p | - | p |
| | Test-taking effort | Reported effort | c | c | p | m of e+v |
| Ability | | Test performance | - | - | c | c |

*Note*. c: criterion, p: predictor, m: mediator, -: not included.

## 5.4 Results

*Measurement model for test-taking motivation scales*

Confirmatory factor analyses were conducted to test the measurement model for all five test-taking motivation constructs. Separate models were estimated for the two time points. The QCM contained twelve items and the test-taking effort scale contained three items. Each factor consisted of three items as described in the method section. All cross-loadings were fixed to zero; correlations between the factors were freely estimated.

Table 5.2

*Factor correlations for the test-taking motivation constructs before and after the test*

| | Effort | Probability of success | Importance | Interest |
|---|---|---|---|---|
| Probability of success | .19* / .26* | | | |
| Importance | .78* / .77* | .03 / .10* | | |
| Interest | .71* / .67* | .19* / .17* | .82* / .79* | |
| Anxiety | .28* / .28* | -.33* / -.31* | .48* / .47* | .28* / .35* |

*Note*. Coefficients refer to the administration of the scales before the test / after the test. $p < .001$.

For this correlated five-factor model, fit was unsatisfactory at both time points (before the test: $\chi^2 = 5211.9$, $df = 80$, $p < .001$; CFI = .93; TLI = .91; RMSEA = .04; SRMR = .05; after the test: $\chi^2 = 4000.7$, $df = 80$, $p < .001$; CFI = .92; TLI = .89; RMSEA = .05; SRMR = .05). Furthermore, the latent factors effort and importance were very closely associated (before the test: $r = .96$; after the test: $r = .97$). A closer inspection revealed that this strong correlation resulted from the fact that one item of the importance scale referred to the effort construct ("I am really going to try as hard as I can on this test"). Consequently, we decided to have this item load on the effort factor instead of the importance factor. In this adapted model, the effort factor was defined by four items and the importance factor by two items. The adapted model showed satisfactory model fit before ($\chi^2 = 3406.7$, $df = 80$, $p < .001$; CFI = .96; TLI = .94; RMSEA = .03; SRMR = .04) as well as after the test ($\chi^2 = 2231.4$, $df = 80$, $p < .001$; CFI = .96; TLI = .94; RMSEA = .03; SRMR = .04). Furthermore, as shown in Table 5.2, the initially very strong correlation between the factors effort and importance decreased. This supported our decision to change the original assignment of the importance item[2].

*Descriptive statistics of the test-taking motivation scales*

As shown in Table 5.3, the reliability estimates of the factors were acceptable both before and after the test given the limited number of items on each scale. The weighted means of the motivation scales before the test suggest that the students planned to invest effort, were confident to be successful and perceived the test as important. Both interest in the test and test anxiety were relatively low. The low level of interest and anxiety is not surprising as it was a low-stakes test with no personal consequences for the test-takers. After the test, the reported invested effort and importance were lower than before the test. Probability of success, interest, and anxiety, in contrast, remained almost stable. In sum, the pattern of student ratings suggested that they were motivated to take the test.

Table 5.3

*Descriptive statistics for the test-taking motivation scales*

| Scale | *N* items | Before the test | | | After the test | | |
|---|---|---|---|---|---|---|---|
| | | *M* | *SD* | ω | *M* | *SD* | ω |
| Effort | 4 | 2.95 | 0.67 | .83 | 2.55 | 0.76 | .85 |
| Probability of success | 3 | 2.88 | 0.56 | .67 | 2.85 | 0.64 | .67 |
| Importance | 2 | 2.71 | 0.80 | .67 | 2.39 | 0.87 | .75 |
| Interest | 3 | 1.99 | 0.63 | .68 | 1.81 | 0.67 | .73 |
| Anxiety | 3 | 1.81 | 0.70 | .72 | 1.70 | 0.70 | .80 |

*Note*. ω. McDonald Omega (McDonald, 1999).

*Prediction of test-taking effort*

Two latent regression models were estimated to predict test-taking effort with perceived value of the test and expectancy for success. First, we were interested in the extent to which test-taking effort is affected by value (research question 1a). Second, we wanted to know if the expectancy component can explain an additional proportion of variance in test-taking effort (research question 1b). Therefore, in the first model we regressed reported effort on importance, interest, and anxiety. In the second model we added probability of success as a fourth predictor. The results for both models and for both time points are given in Table 5.4.

Table 5.4

*Regression of effort on value and expectancy*

| | Model 1a | | | | | | Model 1b | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Before the test | | | After the test | | | Before the test | | | After the test | | |
| | *b* | *(S.E.)* | *β* | *b* | *(S.E.)* | *β* | *B* | *(S.E.)* | *β* | *B* | *(S.E.)* | *β* |
| Importance | 0.40* | (0.02) | .67 | 0.43* | (0.02) | .68 | 0.41* | (0.02) | .68 | 0.43* | (0.02) | .66 |
| Interest | 0.13* | (0.02) | .20 | 0.12* | (0.02) | .17 | 0.10* | (0.02) | .15 | 0.10* | (0.02) | .14 |
| Anxiety | -0.09* | (0.01) | -.09 | -0.10* | (0.01) | -.10 | -0.04 | (0.01) | -.04 | -0.03 | (0.01) | -.03 |
| Probability of success | | | | | | | 0.13* | (0.01) | .13 | 0.17* | (0.02) | .16 |
| $R^2$ *(S.E.)* | .64* | (0.01) | | .61* | (0.01) | | .65* | (0.01) | | .63* | (0.01) | |
| $\Delta R^2$ | | | | | | | .01 | | | .02 | | |
| $\chi^2$ *(df)* | 2035.8 (48)* | | | 1210.6 (48)* | | | 3406.7 (80)* | | | 2231.4 (80)* | | |
| CFI / TLI | .97 / .96 | | | .97 / .96 | | | 96 / .94 | | | .96 / .94 | | |
| RMSEA / SRMR | .03 / .03 | | | .03 / .03 | | | .03 / .04 | | | .03 / .04 | | |

*Note.* Note that Model 1b is a reparameterization of the respective measurement model presented above yielding the same model fit. $\Delta R^2$ refers to increments in explained variance in *effort* by adding *probability of success* as a predictor. * $p < .001$.

In Model 1a all three value aspects (i.e., importance, interest, and anxiety) were significant predictors of test-taking effort and taken together explained over 60% of the variance in effort scores at both time points. Importance was the strongest predictor of effort. Anxiety had a small negative effect. Adding probability of success as a predictor (Model 1b) reduced the negative effect of the anxiety scale which was no longer a significant predictor at either time point. The regression coefficients of importance and interest remained nearly stable. Probability of success had roughly the same standardized regression coefficient as interest, but the amount of explained variance in effort scores increased only by 2% at both time points. Focusing on the sizes of the standardized coefficients, only importance showed a substantial effect (exceeding ± .20) on effort. Thus, the more the students perceived the test as important, the higher was their test-taking effort. To summarize, in line with our assumptions both expectancy and values were related to effort. However, contrary to our expectations, the value aspect importance was by far the most powerful predictor whereas expectancy for success showed only very little incremental validity in predicting effort over and above the value aspects.

*Prediction of test performance*

In a subsequent model, we predicted test performance with test-taking motivation. In Model 2a, effort was the only predictor of test performance and explained just under 10% of the variance in test performance at both time points (before the test $R^2 = .08$, after the test $R^2 = .09$). The standardized regression coefficient was .27 before the test and .31 after the test. Thus, as hypothesized, test-taking effort had a substantial impact on test performance.

Finally, we tested whether the effects of expectancy for success and perceived value of the test on test performance were mediated by test-taking effort. As shown in Figure 5.2, for both time points the regression models had satisfactory fit. The prediction of test-taking effort with the three value scales and probability of success showed the same effects as described for Model 1b: Importance was the strongest predictor of effort and explained over 60% of the variance in effort scores at both time points.

Taking *direct effects* of the value scales and probability of success into account, the amount of variance in test performance explained by all test-taking motivation constructs including effort increased to 18% and 28% before and after the test, respectively. Before the test, significant predictors of test performance were reported effort and probability of

success; after the test, interest also had a significant but relatively small and negative effect. Thus, the higher the students' self-reported probability of success and effort, the better they scored on the test. The predictive power of probability of success increased from before to after the test, while the predictive power of effort remained stable. These results are in line with our hypothesis that expectancy for success explained additional variation in mathematics ability estimates beyond test-taking effort.

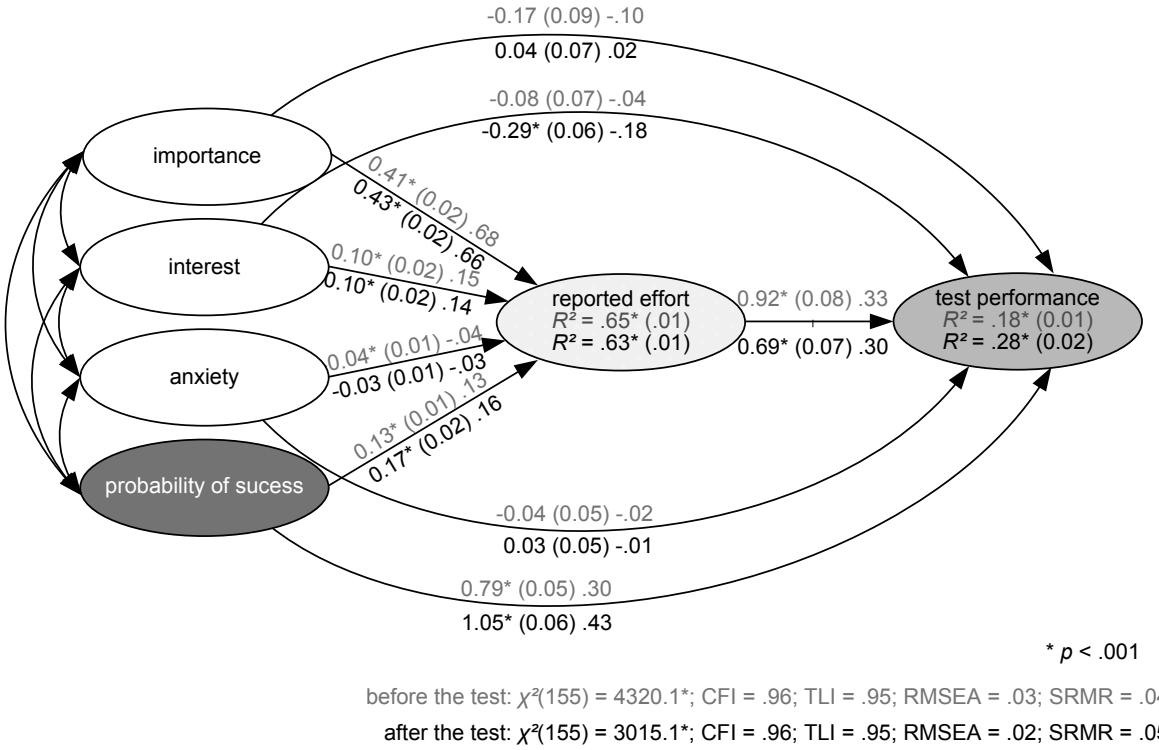



*Figure 5.2*. Model 2b, Regression of test performance on expectancy for success and value mediated by test-taking effort (Coefficients designate *b (SE) β*. Grey font: Before the test. Black font: After the test. Manifest indicators and disturbance terms omitted for simplicity).

Regarding the *indirect effects* of the value and expectancy components, it is apparent that importance had the strongest indirect effect on test performance via effort ($b = 0.38$, $SE = 0.04$, $p < .001$, $\beta = .23$; *not illustrated*). Due to the non-significant negative direct effect of importance on test performance, we concluded that importance was almost completely mediated by test-taking effort. The indirect effects of probability of success ($b = 0.12$, $SE = 0.01$, $p < .001$, $\beta = .04$; *not illustrated*) and interest ($b = 0.09$, $SE = 0.02$, $p < .001$, $\beta = .05$; *not illustrated*) were significant but rather small. The indirect effect of anxiety did not reach significance. After the test, the indirect effects of the value scales and

probability of success on test performance remained nearly stable. Again, the indirect effect of importance on test performance via effort was the strongest one ($b = 0.30$, $SE = 0.03$, $p < .001$, $\beta = .20$; *not illustrated*) and the direct effect of importance on test performance was not significant. Probability of success ($b = 0.12$, $SE = 0.02$, $p < .001$, $\beta = .05$; *not illustrated*) and interest ($b = 0.07$, $SE = 0.01$, $p < .001$, $\beta = .04$; *not illustrated*) showed a fairly small but significant indirect effect on test performance. Thus, the more important the students perceived the test, the more effort they invested; in turn, the more effort they invested and the higher they perceived their probability of success, the better they performed on the test.

In sum, the results of the latent regression analyses suggested that the effect of the value component on test performance was partially mediated by test-taking effort. Importance was the strongest predictor of reported effort and showed the strongest indirect effect on test performance. Probability of success also had a pronounced direct effect on test performance that increased after the test, whereas the effect of effort remained stable. Altogether, more than one quarter of the variance in mathematics ability could be explained with this mediation model based on test-taking motivation measurement after the test.

## 5.5 Discussion

The current study investigated the complex construct of test-taking motivation and its relationship to test performance in a representative sample of more than 40,000 ninth-graders. On the basis of expectancy-value theory (Wigfield & Eccles, 2000) we examined the relationships among expectancy for success, perceived value of the test, test-taking effort, and test performance. Other studies grounded on expectancy-value theory mostly assessed only two aspects of test-taking motivation, such as effort and value or expectancy and value (Asseburg, 2011; Cole et al., 2008; Eklöf et al., 2014). Thus, it was important to shed light on the interrelations among all components as well as their association with performance. We pursued four main objectives: 1) establishing a measurement model of the five test-taking motivation scales administered in the study; 2) predicting test-taking effort with the expectancy and value scales; and 3) evaluating a mediator model in which the effects of the expectancy and value scales on test performance were partially mediated by test-taking effort. Since previous studies that assessed test-taking motivation before the test obtained different results than studies that assessed test-taking motivation after the test, we were also interested in a comparison of the effects for test-taking motivation before the

test versus test-taking motivation after the test. Thus, we administered the motivation questionnaire both before and after the test.

*Measurement model for the test-taking motivation scales*

Before conducting regression analyses, we established a measurement model of the five test-taking motivation scales for the time points before and after the test. This was necessary because the motivational scales had not been administered together in previous research. The final model showed satisfactory fit for both time points. Correlations between the scales indicated that importance, interest, and effort are highly correlated confirming the theoretical relationship between value and test-taking effort (Wigfield & Eccles, 2000). In sum, our questionnaire was apparently suitable for assessing test-taking motivation both before and after the test. However, only two items were used for the assessment of importance which resulted in limited content coverage. In future research, additional items should be used to assess this facet.

*Prediction of test-taking effort*

In subsequent analyses, we predicted test-taking effort using the scales for the value and expectancy components. At both time points, the scales for the value component explained over 60% of the variance in effort scores. The addition of probability of success (i.e., the expectancy component) increased the proportion of explained variance only slightly. Hence, the value component was more important for the prediction of effort than the expectancy component.

In our study, importance was the main predictor of test-taking effort. We measured importance indirectly via the challenge scale of the QCM. Challenge refers to the extent to which test-takers view the situation as an achievement situation. If the test-takers seek success, the test becomes more important, which corresponds with attainment value in the expectancy-value model (Vollmeyer & Rheinberg, 2006). Other studies also revealed effects for the various value aspects such as usefulness and importance of the test (Cole et al., 2008), test attractiveness (Penk, Pöhlmann, & Roppelt, 2014), or in terms of the expectancy-value theory, the utility, attainment, and intrinsic value, respectively. Thus, the results from the present study are in line with current research. However, our measures were much more powerful predictors of effort scores (in terms of variance explained) than those in previous studies (Cole et al., 2008; Penk et al., 2014). It is possible that in a low-stakes context, measures that serve as subtle assessors of the attainment value (i.e.,

measuring challenge) are better predictors of test-taking effort than more direct questions about the importance of the test.

*Prediction of test performance*

Our next research question pertained to the prediction of test performance. First, we used test-taking effort as a predictor and could explain almost 10% of the variance in mathematics ability at both time points. These results are in line with previous research that found similar effects of test-taking effort on test performance (Baumert & Demmrich, 2001; Eklöf & Nyroos, 2013; Eklöf et al., 2014).

In the mediating model with expectancy, value, and effort it became apparent that expectancy and effort had strong direct effects on test performance. Based on these findings, the value component seems to be more important for the prediction of effort than for the prediction of test scores. In our study we were able to explain more than a quarter of the variance in mathematics scores by differences in test-taking motivation measured after the test. The same proportion was found in the study of Baumert and Demmrich (2001) who showed that the effort and worry scales served as mediator variables for the other constructs. This underlines the importance of test-taking motivation for the prediction of effort and test performance.

Our results partially support the findings of Freund and Holling (2011) who demonstrated probability of success to be a strong predictor of test performance (as in our study) besides interest (as not in our study). However, our findings contrast with the results of Freund et al. (2011) who used the QCM and found only interest to be a significant predictor of test performance. Surprisingly, in our study, interest measured after the test showed a significant negative relationship with test performance, but the standardized regression coefficient did not exceed an absolute value of .20. The negative coefficient found for interest may be due to multicollinearity among the predictors; on the other hand, the correlations between the predictors were mostly low to moderate. As expected for low-stakes tests and as reflected by the mean of the interest scale, this test was not especially appealing for the students. Moreover, the bivariate correlation between interest measured after the test and test performance was indeed positive but close to zero. Thus, it seems that in a low-stakes context, interest in the test plays a minor role for the prediction of test performance but a more important role for the prediction of test-taking effort.

The current study makes an important contribution to the field by demonstrating that expectancy for success is relevant for the prediction of test performance in low-stakes assessments. This implies that researchers investigating test-taking motivation and applying the expectancy-value model should include measures of the expectancy component in addition to test-taking effort and the value component. Further research should investigate whether the influence of probability of success persists if domain-specific competence beliefs are added, such as self-concept in the relevant subject. Moreover, an implication for testing considers the equivalence of ability level of the students and difficulty of the test. From a psychometric point of view, there is nothing new in this notion. However, a match between the ability level of the test-taker and test difficulty is also desirable from a motivational point of view. In fact, an optimal level of test-taking motivation may be achieved if the difficulty of the test is lower than the test-taker's ability level. If the test is extremely difficult relative to student ability, it is likely that poor performance is also due to a lack of effort (Asseburg & Frey, 2013). Thus, a low probability of success could be an indication of an overly arduous test. In fact, Asseburg and Frey (2013) recommended a mean success probability of 70% because test-takers invested more effort and reported less boredom and daydreaming the easier the items were. However, it is impossible to construct a single test that perfectly matches the ability levels of all students. As a consequence, even for a test with a desirable mean success probability, the test will be much more difficult (i.e., possibly less motivating) for some test-takers than for others. This issue is addressed by computerized adaptive testing (CAT). In such tests the items are individually selected as a function of the responses on previous items. Thus, in CAT the test is tailored to the ability level of the test-taker (Frey & Seitz, 2011) to achieve an optimal fit of test difficulty and student ability. However, CAT poses specific organizational and methodological challenges, especially in a large-scale assessment context (Wainer, 2000).

On the basis of Cole et al. (2008), we also tested whether test-taking effort represented a mediator variable for the effect of expectancy and value on test performance. Our hypothesis that the effects of both the expectancy and the value component on test performance are mediated by effort received only partial support: In the current study, the effect of the expectancy component on test performance was not mediated by test-taking effort. However, the results showed that at both time points, the effect of importance (attainment value) on test performance was mediated by test-taking effort. This finding is

in line with previous research that found full mediation for usefulness (Cole et al., 2008) and importance of the test (Cole et al., 2008; Zilberberg et al., 2014), two aspects of the value component. Therefore, emphasizing the importance of the test could be a means of increasing the perceived value of the test, and in turn, students' efforts. Lau and colleagues (2009) presented proctor strategies to enhance students' test-taking motivation. They found that the behavior of the proctors (e.g., emphasizing the importance and usefulness of the test; encouraging test-takers to give their full effort during the testing session) affected students' invested effort. Furthermore, Zilberberg and colleagues (2014) found that conveying the purpose of the test affected the perceived importance of the test and suggested strategies such as "well-crafted test instructions, speeches from respected individuals, and educational videos" (Zilberberg et al., 2014, p. 377) to promote the importance of a test. In our view, the importance and usefulness of the test should at least be emphasized in the test instructions.

Probability of success and interest had only small indirect effects on test performance via effort, but had stronger direct effects. Anxiety had no major influence on either effort or test performance. Effort and test anxiety were also uncorrelated in the Swedish National Test (Eklöf & Nyroos, 2013). Thus, in low-stakes testing situations, anxiety does not seem to be very important for either effort or test performance. Our findings also correspond with a study in the context of achievement motivation that found a mediating role of effort on the relationship between motivational regulation strategies and achievement (Schwinger, Steinmayr, & Spinath, 2009).

Additionally, the results show that the predictive power of the expectancy component increased from before to after the test, whereas the predictive power of the value component remained stable. Thus, the results of Freund et al. (2011) who found only interest to be a significant predictor of performance might be explained by their pre-test use of the QCM. Another indication in support of this interpretation is the study of Freund and Holling (2011). In their study, the application of the QCM in a retest revealed increased effects of all the QCM scales for the prediction of test performance. The students were familiar with the test when the motivation questionnaire was used before the retest. This is similar to our assessment of test-taking motivation after the test in the current study. In our study, the strongest predictors of test performance were probability of success and test-taking effort. Thus, our results are in line with previous research in that the value component played a minor part in predicting test performance (Schunk et al., 2008).

Furthermore, the results showed that initial test-taking motivation had less predictive power than test-taking motivation after the test. The relationship between test-taking motivation – especially probability of success – and performance was stronger after the test, which indicates that the assessment of motivation may be more accurate after the administration of the test. It is possible that probability of success assessed before the test represents something different than probability of success assessed after the test. However, it is perhaps not surprising that students provide more precise indications of their test-taking motivation after working on the test. Probability of success should correspond with the perceived difficulty of the test so that students' test-taking motivation should change depending on the items they worked on (Vollmeyer & Rheinberg, 2006). On the other hand, students' initial motivation creates a respective emotional state in the test-takers which corresponds to the amount of effort they are willing to invest in the test (Boekaerts, 2007).

*Limitations and conclusions*

Several limitations should be noted. First, we modified the original factor definitions with respect to the scales importance and effort. Due to this adaption, only two items constituted the importance factor. Generally, more items per scale would have been desirable. Another point concerns the various aspects of the value component. In the current study, we covered the intrinsic value with interest, the cost aspect with anxiety, and the attainment value only indirectly with importance. It is possible that assessing the attainment value directly with perceived importance of the test would lead to other results. Specifically, it is conceivable that perceived importance would have had a smaller effect on test performance than challenge because the students do not perceive the test as important but felt challenged nonetheless. Furthermore, we did not measure the utility value. Further research should analyze the relationship of this value aspect with test-taking effort and test performance. This is especially interesting for low-stakes tests. Such tests without individual feedback for the test-takers are especially useful on the system level for the evaluation of the quality of the schooling system but are of rather limited usefulness for the individual test-takers.

Besides the expectancy and value components, other motivational constructs are considered in the expectancy value model such as perception of competence and affective memory as well as variables concerning cognitive processes and the social world (for a detailed description of the whole model see Wigfield & Eccles, 2000). Consideration of the

complete model was beyond the scope of this article. Of course, this does not mean that the two main components are not influenced by the other factors in the overall model. Here, future research is merited.

Finally, recent research found empirical support for the assumed interaction between expectancy and value in achievement motivation settings (Nagengast et al., 2011; Trautwein et al., 2012). Further studies should expand the model proposed in this study to explore possible interaction terms in the test-taking motivation context in line with Atkinson's theory of achievement motivation (1957).

In conclusion, to accurately model the entire test-taking motivation construct, all three components – expectancy for success, perceived value of the test, and test-taking effort – should be considered when evaluating the effects of motivation on test scores. Otherwise, one risks missing an important aspect in understanding motivational influences on test performance.

## Notes

[1] As a prerequisite for the regression analyses, we checked for multicollinearity by calculating the Variance Inflation Factor (VIF) for all motivational constructs including effort based on the latent variables in the final measurement model. VIF values for the measurement before the test varied between 1.29 (probability of success) and 5.08 (importance). VIFs for the measurement after the test were almost identical or slightly lower. In sum, VIFs did not indicate substantial multicollinearity according to common cut-off values (see e.g., Cohen, Cohen, West, & Aiken, 2003).

[2] The assumption of measurement invariance of the five constructs is plausible due to the short time difference of approximately two hours between the two time points. Penk and Richter (2015) supported strong longitudinal measurement variance for the scales reflecting probability of success, importance, and effort.

## References

Asseburg, R. (2011). *Motivation zur Testbearbeitung in adaptiven und nicht-adaptiven Leistungstests [Test-taking motivation in adaptive and sequential achievement testing]* (Doctoral dissertation). Christian-Albrechts-Universität zu Kiel. Retrieved from the website http://d-nb.info/1013153863/34

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling, 55*(1), 92–104.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review, 64*(6), 359–372.

Atkinson, J. W. (1964). *An introduction to motivation*. New York: Van Nostrand.

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education, 16*(3), 441–462.

Boekaerts, M. (2007). What have we learned about the link between motivation and learning/performance? *Zeitschrift Für Pädagogische Psychologie, 21*(3), 263–269. doi:10.1024/1010-0652.21.3.263

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology, 33*(4), 609–624.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah: LEA.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology, 53*(1), 109–132. doi:10.1146/annurev.psych.53.100901.135153

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing, 7*(3), 311–326.

Eklöf, H. (2010). *Student motivation and effort in the Swedish TIMSS Advanced field study*. Paper presented at the 4th IEA International Research Conference, Gothenburg.

Eklöf, H., & Nyroos, M. (2013). Pupil perceptions of national tests in science: Perceived importance, invested effort, and test anxiety. *European Journal of Psychology of Education, 28*(2), 497–510. doi:10.1007/s10212-012-0125-6

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS advanced. *Applied Measurement in Education, 27*(1), 31–45. doi:10.1080/08957347.2013.853070

Enders, C. K. (2010). *Applied missing data analysis*. New York: Guilford Press.

Freund, P. A., & Holling, H. (2011). Who wants to take an intelligence test? Personality and achievement motivation in the context of ability testing. *Personality and Individual Differences, 50*(5), 723–728. doi:10.1016/j.paid.2010.12.025

Freund, P. A., Kuhn, J. T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences, 51*(5), 629–634. doi:10.1016/j.paid.2011.05.033

Frey, A., & Seitz, N.-N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in the Programme for International Student Assessment. *Educational and Psychological Measurement, 71*(3), 503–522. doi:10.1177/0013164410381521

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice, 28*(3), 39–53. doi:10.1111/j.1745-3992.2009.00154.x

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55.

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education, 58*(3), 196–217. doi:10.1353/jge.0.0045

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, N.J: L. Erlbaum Associates.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide. Seventh edition*. Los Angeles, CA: Muthén & Muthén.

Nagengast, B., Marsh, H. W., Scalas, L. F., Xu, M. K., Hau, K.-T., & Trautwein, U. (2011). Who took the "x" out of expectancy-value theory? A psychological mystery, a substantive-methodological synergy, and a cross-national generalization. *Psychological Science, 22*(8), 1058–1066. doi:10.1177/0956797611415540

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (2013). *The IQB National Assessment Study 2012. Competencies in mathematics and the sciences at the end of secondary level I. Summary*. Münster: Waxmann.

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: An investigation of school-track-specific differences. *Large-scale Assessments in Education, 2*(1). doi:10.1186/s40536-014-0005-4

Penk, C., & Richter, D. (2015). Change in test-taking motivation and its relationship to test performance in large-scale assessments. *Manuscript Submitted for Publication.*

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern-und Leistungssituationen [A questionnaire for the measurement of current achievement motivation in learning and achievement situations]. *Diagnostica, 47*(2), 57–66.

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). *Motivation in education: Theory, research, and applications* (3rd ed.). Upper Saddle River (NJ): Pearson Education.

Schwinger, M., Steinmayr, R., & Spinath, B. (2009). How do motivational regulation strategies affect achievement: Mediated by effort management and moderated by intelligence. *Learning and Individual Differences, 19*(4), 621–627. doi:10.1016/j.lindif.2009.08.006

Stanat, P., & Lüdtke, O. (2013). International large-scale assessment studies of student achievement. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 481–483). New York, NY: Routledge.

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education, 24*(2), 162–188. doi:10.1080/08957347.2011.555217

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. *The Journal of General Education, 58*(3), 129–151. doi:10.1353/jge.0.0047

Trautwein, U., Marsh, H. W., Nagengast, B., Lüdtke, O., Nagy, G., & Jonkmann, K. (2012). Probing for the multiplicative term in modern expectancy–value theory: A latent interaction modeling study. *Journal of Educational Psychology, 104*(3), 763–777. doi:10.1037/a0027470

Vollmeyer, R., & Rheinberg, F. (2006). Motivational effects on self-regulated learning with different tasks. *Educational Psychology Review, 18*(3), 239–253. doi:10.1007/s10648-006-9017-0

Wainer, H. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah, N.J: Lawrence Erlbaum Associates.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*(3), 427–450.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review, 6*(1), 49–78. doi:10.1007/BF02209024

Wigfield, A., & Eccles, J. S. (2000). Expectancy–Value theory of achievement motivation. *Contemporary Educational Psychology, 25*(1), 68–81. doi:10.1006/ceps.1999.1015

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment, 10*(1), 1–17. doi:10.1207/s15326977ea1001_1

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*(3), 227–242. doi:10.1207/s15324818ame0803_3

Zilberberg, A., Finney, S. J., Marsh, K. R., & Anderson, R. D. (2014). The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model. *International Journal of Testing, 14*(4), 360–384. doi:10.1080/15305058.2014.928301

# 6

# Studie III

## *Change in Test-Taking Motivation and its Relationship with Test Performance in Low-Stakes Assessments*

Penk, C. & Richter, D. (2015). Change in Test-Taking Motivation and its Relationship with Test Performance in Low-Stakes Assessments. *Manuscript submitted for Publication*.

(Stand: Februar 2015)

## Abstract

Since the turn of the century, an increasing number of low-stakes assessments (i.e., assessments without direct consequences for the test-takers) are being used to evaluate the quality of educational systems. Internationally, research has shown that low-stakes test results can be biased due to students' low test-taking motivation, and that students' effort levels can vary throughout a testing session involving both cognitive and noncognitive tests. Thus, it is possible that students' motivation vary throughout a single cognitive test and in turn affect test performance. This study examines the change in test-taking motivation within a two-hour cognitive low-stakes test and its association with test performance. Based on expectancy-value theory, we assessed the three components of test-taking motivation: expectancy, value, and effort. Thus, besides the change in effort we also modeled the change in perceived value of the test and expectancy for success an often ignored component in studies applying the expectancy-value theory. Using data from a large-scale educational assessment study of German ninth-graders, we employed second-order latent growth modeling and structural equation modeling to predict test performance in mathematics. On average, students' effort and perceived value of the test decreased, whereas expectancy for success remained stable. Overall, initial test-taking motivation was a better predictor of test performance than change in motivation. Only the variability of change in the expectancy component was positively related to test performance. The theoretical and practical implications for test practitioners are discussed.

*Keywords: test-taking motivation, low-stakes tests, large-scale assessments, expectancy-value theory, growth modeling*

# Change in Test-Taking Motivation and its Relationship with Test Performance in Low-Stakes Assessments

## 6.1 Introduction

Students approach tests with different attitudes and affects and thus may engage in different behaviors. Tests with short- or long-term consequences for the students are called *high-stakes* tests. Given the inferences made from test scores, testing practitioners seem to be assuming that students give high effort during high-stakes tests due to the personal consequences for the test-takers. However, since the beginning of the Program for International Student Assessment (PISA) in the year 2000, assessments without consequences for the test-takers (i.e., *low-stakes* tests) have become increasingly important for evaluating the quality of Germany's educational system (Stanat & Lüdtke, 2013). Test-taking motivation (TTM) is an important issue under these circumstances, because it is possible that students do not give their best effort due to the lack of any personal consequences of the test results.

TTM refers to students' readiness to engage in completing the test (Baumert & Demmrich, 2001). Research in this field has increased in the past decade, and several studies have shown that TTM can affect test performance (Baumert & Demmrich, 2001; Cole, Bergin, & Whittaker, 2008; Eklöf, Pavešič, & Grønmo, 2013; Thelk, Sundre, Horst, & Finney, 2009). A synthesis of twelve empirical studies showed that motivated students outperformed their unmotivated classmates by more than half a standard deviation in low-stakes tests (Wise & DeMars, 2005). Thus, the impact of motivation on performance is very important for the interpretation of test results in low-stakes assessments. If the students do not give their best effort, it remains unclear whether test scores correspond to the true ability of the students. As noted by Barry, Horst, Finney, Brown, and Kopp (2010), both the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) and the Guidelines on Test Use from the International Testing Commission (International Test Commission, 2001) recommend collecting information associated with student motivation.

In addition to the issue of initial motivation with which students approach the testing session, it is also important to consider the potential *change* in motivation throughout a long testing session and its effect on students' test performance. TTM can both increase and decrease throughout a test. An increase in TTM can be interpreted as flow (i.e., a state

with a challenge-skill balance in which the individual feel at the same time cognitively efficient, motivated, and happy; Moneta & Csikszentmihalyi, 1996). A decrease in TTM can be interpreted as fatigue or boredom. However, in a low-stakes testing context it is more conceivable that the students show a decrease in motivation within a cognitive (mentally taxing) test. A study investigating the change in motivation found no fatigue effect (i.e., a decline in effort over the course of the testing session) during a low-stakes testing session with different test types, i.e., cognitive and noncognitive test (e.g., measures of attitudes). However, it did find that TTM was influenced by test-specific characteristics, such as mental taxation (Barry & Finney, in press). Studies exploring the change in TTM within one cognitive test found a decrease in TTM (Horst, 2010; Wise, 2006; Wise, Pastor, & Kong, 2009), but they did not investigate the relationship between the change in TTM during the cognitive test and test performance. This relationship, however, is of particular interest for large-scale assessments evaluating the outcomes of educational systems. Using the results of low-stakes tests without knowing the effect of change in TTM on test results can threaten the validity of inferences based on those test results (Eklöf, 2008, 2010a; Thelk et al., 2009). Thus, the current study aims to investigate (a) the change in TTM based on the expectancy-value theory (EVT) and (b) the relationship between change in TTM and test performance in a German low-stakes large-scale assessment. This investigation is especially important for several reasons: a) change in effort during a test is often discussed but rarely empirically evaluated in research dealing with TTM; moreover, when it is studied, change in effort is not linked to actual test performance; b) the expectancy component is often ignored in applications of the expectancy-value theory; and c) the role of effort as a mediator variable between performance and both the expectancy and value components requires further investigation. Each of these issues is explained in more detail below. In the next section, we first define the construct of TTM and integrate it into the framework of EVT. We then provide an overview of previous research.

*Expectancy-Value Theory and the Non-Longitudinal Assessment of TTM*

EVT is a frequently used framework in the context of TTM (Eccles & Wigfield, 2002; Sundre, 2007; Wigfield & Eccles, 2000). As shown in Figure 6.1, EVT assumes that the expectancies for success and the perceived value of a test directly affect achievement behavior, which involves both the expended effort on the test and actual test performance. Expectancies refer to students' perceptions of how well they will perform, and therefore include each individual's perception of his or her own competence in a given task.

Students may ask themselves, "*Can I do well on this test?*" The value component includes four distinct aspects: attainment value (the importance of the test), intrinsic value (the enjoyment during the test), utility value (the usefulness of the test), and cost (test anxiety or expended effort). The value component essentially concerns the question, "*Why should I take this test?*"

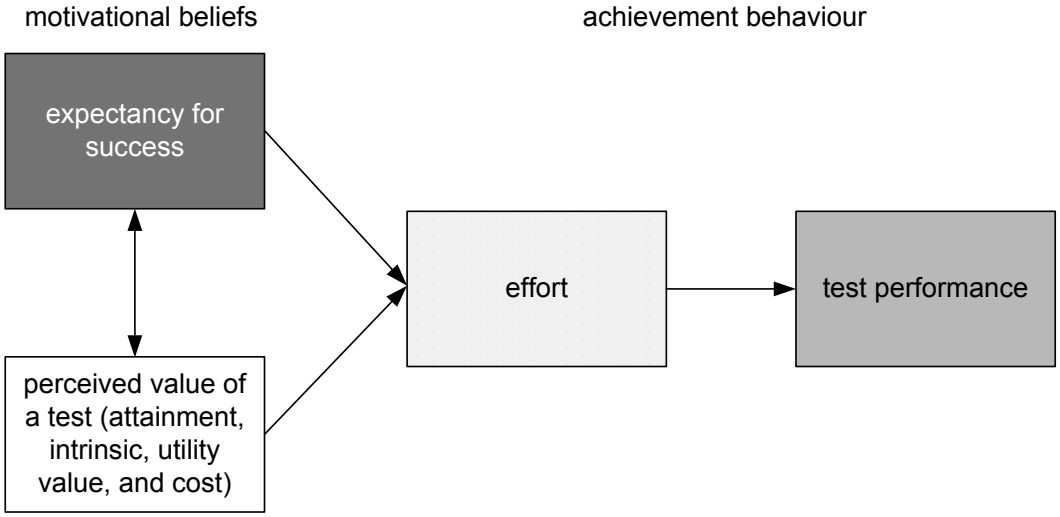



*Figure 6.1*. Expectancy-value theory in the context of test-taking motivation (adapted from Eccles & Wigfield, 2002; Wigfield & Eccles, 2000).

Test-taking motivation is defined as "the willingness to engage in working on test items and to invest effort and persistence in this undertaking" (Baumert & Demmrich, 2001, p. 441). Thus, test-taking effort constitutes the main element of TTM and is described as the engagement of the test-takers and their expenditure of energy to achieve the best possible test score (Wise & DeMars, 2005). According to EVT, effort is the outcome of expectancy and value, and is therefore related to test performance. This means that effort should mediate the relationship between performance on the one hand and expectancy and value on the other hand. In light of EVT, the TTM construct includes all three components: the effort that the students invest, their expectancy for success, and the value they place in the test.

Typically in school settings, the expectancy component has been shown to be a stronger predictor of test performance than the value component, which was more closely associated with persistence or task choice (Eccles & Wigfield, 2002; Pekrun, Elliot, & Maier, 2009; Schunk, Pintrich, & Meece, 2008; Wigfield, 1994; Wigfield & Eccles, 2000). However, in low-stakes tests, most research only considers the value component and effort

when examining TTM (Cole et al., 2008; Eklöf & Nyroos, 2013; Eklöf et al., 2013; Wolf & Smith, 1995) and ignores the expectancy component, "because test-takers in low stakes tests seldom have any way of finding out if they were successful" (Cole et al., 2008, p. 613). Overall, studies have found a positive relationship between the value component and test performance, as well as between effort and test performance. Cole et al. (2008) have investigated and found support for the mediating role of effort. Specifically, they found partial mediation of the intrinsic value on test performance and full mediation of the attainment and utility value on test performance. Zilberberg et al. (2014) found effort fully mediated the effect of attainment value on performance. The few studies that included both the expectancy and value components found that expectancy for success predicted test performance in low-stakes tests (Asseburg, 2011; Freund & Holling, 2011); however, these studies did not examine effort. In sum, most of the studies ignored at least one aspect of TTM or the mediating role of test-taking effort.

*Longitudinal Assessment of TTM*

The longitudinal assessment (i.e., the change in TTM) of motivation is important, because achievement tests often take much longer than a regular school period (45 minutes). Therefore, a long testing session may lead to fatigue and a decrease in TTM (Cao & Stokes, 2008). According to EVT, it is also conceivable that a loss of high expectancy for success during the test can result in a decline in effort, resulting in low test performance. In particular, due to recurring difficult items that the test-takers cannot solve might change their confidence to answer the next items and their willingness to invest effort. This dynamic was uncovered by Wise and Smith (2011) in their demands-capacity model of test-taking effort that includes aspects of initial effort and potential change in effort during the test. They assumed that an test-takers' effort behavior is determined by two effort constructs: (a) the resource demands of a specific test item (e.g., item difficulty), which can be thought of as the necessary effort to solve the item, and (b) the effort capacity of the test-taker, which refers to the amount of effort the test-taker willing to expend on the item. The effort capacity includes the determinants of the initial effort capacity (e.g., test consequences) as well as the determinants of the effort capacity from previous items on the test (e.g., the amount of fatigue or change in confidence to solve future items), which can vary between test-takers and within test-takers during the course of the test. The authors also emphasized the dynamics of TTM to make a meaningful interpretation of the test results. However, very few studies consider the dynamic of TTM in their analyses.

Some studies have examined these dynamics or the change in TTM over a low-stakes testing session using response behavior at the item level in computer based assessments (Wise, 2006; Wise et al., 2009). The assessments used in those studies recorded the time the test-takers spent to complete an item, and based on this response time, the behavior was classified as solution behavior (effortful responses) or rapid-guessing behavior (non-effortful responses). This classification indirectly mirrors the motivation of the test-taker, because one can assume that unmotivated test-takers do not spend much time reading and answering the items on the test. *Response time fidelity* reflects the proportion of test-takers who have shown solution behavior on an item (Wise & Kong, 2005; Wise & Smith, 2011). Wise (2006) and Wise et al. (2009) found a negative relationship between item position and response time fidelity scores; if the item was placed near the end of the test, test-takers were less likely to "solve" the item. This indicates that TTM may decrease throughout the test, for instance, due to fatigue effects or change in confidence in the ability to answer future items correctly (Wise & Smith, 2011). At present, however, most large-scale assessments are traditional paper-and-pencil tests and one cannot use an electronically recorded measure of TTM. For this type of assessment one can only employ self-reported measures of motivation. Although self-reported measures are not ideal, it has been shown that they could also serve as valid indicators of test-taking motivation in low-stakes tests (Swerdzewski, Harmes, & Finney, 2011).

Thus, other studies have examined change in TTM on paper-and-pencil tests via self-report measures of motivation. These studies focused on three hour long testing sessions including one cognitive and four noncognitive tests (Barry & Finney, in press; Barry et al., 2010; Horst, 2010). For the assessment of TTM they used the *Student Opinion Scale* (SOS; Sundre, 2007) after completing each of the five tests, which includes two components: perceived importance of the test (attainment value) and test-taking effort. Overall, the studies found no indication of a fatigue effect within the testing session, but they showed that effort was tied to the type of the assessment. Specifically, in this low-stakes testing session students reported less effort on the cognitive test and more effort on the noncognitive tests (Barry & Finney, in press; Barry et al., 2010). These effects were independent of the order in which the tests were administered what suggested that in low-stakes assessments students are less willing to invest effort in mentally taxing tests. Moreover, Barry and Finney (in press) investigated the *change* in effort and importance across one cognitive and four noncognitive tests. In general, effort slightly increased during the testing session,

with the smallest reported effort score found for the cognitive test that was administered first. Barry and Finney assumed that the rise in effort over the testing session is probably due to the low mental taxation of the noncognitive test in relation to the high mental taxation of the cognitive test (i.e., higher cost in light of EVT). Thus, students' effort may decrease within one high mental taxing test. In contrast, students rated the cognitive test as the most important, even though they invested the least effort in it in comparison to the other tests. Effort and importance were moderately correlated for the cognitive test, but the *change* in effort and the *change* in importance were not. Especially for the cognitive test (with high mental taxation) it seems important that the test-takers consider the test as important so that they invest effort. Although in this study the change in effort and change in importance in within all five tests was not related, it is possible that the change in effort and change in importance within a cognitive test is related. Barry and Finney also assessed the expectancy component (i.e., self-efficacy in mathematics) after the students completed all tests; however, it was not related to students' effort on either the cognitive or noncognitive tests (Barry & Finney, in press). Importantly, this measure was collected only once for the cognitive test, so no conclusion about the change in expectancy can be drawn. Self-efficacy in mathematics was used as a domain-specific measure of expectancy, but a situation-specific measure of expectancy for success was lacking. It is conceivable that the expectancy for success for a specific test differs from the general expectancy for success in a domain.

Horst (2010) used a design similar to that of Barry and Finney (in press). She split the cognitive test administered at the beginning of the assessment into three subtests to assess how TTM changes *within a cognitive test*. Students showed a slight decrease in effort within the cognitive test, most likely indicating a weak fatigue effect. In contrast, the reported level of perceived importance of the test was stable throughout the entire session. Although the test was viewed as important the entire time, the reported effort decreased, probably due to the high "cost" of the cognitive test. It is possible that the student demonstrated a lower level of test performance due to the diminished level of effort than they could have demonstrated with a stable level of effort. Additionally, effort and importance showed higher relationships for the cognitive test than for the noncognitive tests, indicating that the two components of EVT are more closely associated for cognitive tests. Horst recommended administering test-taking effort items also prior to the tests to collect a measure of initial TTM. This could indicate whether some students are unmotivated from

the beginning of the testing session or whether TTM changes throughout the testing session, possibly due to the influence of the completion of previous test items.

A recent study examined the relationship between expectancy, value, test-taking effort, and test performance (Penk & Schipolowski, 2014). This study integrated all three measures of the full TTM construct (expectancy, value, and test-taking effort) in a large-scale low-stakes testing context. In contrast to most studies assessing TTM after a test, students completed a TTM questionnaire before and after a two-hour cognitive test. At both time points, test-taking effort was predicted by the value component of EVT and test-taking effort as well as expectancy for success were strong predictors of test performance. The effect of perceived importance of the test (attainment value in EVT) on test performance was fully mediated by effort before and after the test. Interestingly, the relationship between the expectancy component and test performance was stronger for the measure of expectancy collected *after* the test ($\beta = .43$) than before the test ($\beta = .30$). Perhaps after the test students have a better idea of the difficulty of the test and of how they did on the test. Thus, the often neglected expectancy component seems to be an important construct in the study of TTM in low-stakes assessments. The relationship between effort and test performance was equivalent for both measurement points. Although the authors examined all aspects of TTM, they do not provide any information on the change of motivation during the test. Examining change in the construct is important because students may demonstrate different motivational trajectories within a test or a testing session that can influence test performance.

To summarize, only a few studies (Horst, 2010; Wise, 2006; Wise et al., 2009) have investigated the change in TTM within a cognitive test; they found evidence for a small fatigue effect. Additionally, only one study (Barry & Finney, in press) has assessed the expectancy component; however, this study did not assess the change in expectancy for success throughout the testing session. Furthermore, no study has investigated the relationship between change in TTM within a cognitive test and test performance. This is especially important for large-scale assessments such as PISA, because changes in TTM may impact the results of the tests and therefore limit or even bias their interpretation. Test scores from test-takers who show a decrease in TTM are likely to underestimate actual ability or achievement.

## 6.2 Study Objectives and Research Questions

The purpose of this study is to build upon previous research (Barry & Finney, in press; Horst, 2010; Penk & Schipolowski, 2014) by modeling the *change in all three components of EVT* (expectancy, value, and effort) and relating the change in these constructs to test performance. TTM was assessed within a two-hour low-stakes assessment at three measurement points: before the test, after half of the test, and after the test. Thus, this study is similar to the one conducted by Horst (2010) in that TTM was also assessed throughout a cognitive test. Unlike to Horst (2010), however, this study also assessed initial TTM measured before the test and change in the expectancy component. Altogether, we focus on two types of change: a) average change and b) variability in intra-individual change. The average change (a) describes the mean rate of change of *all* students within the testing session. The variability in intra-individual change (b) refers to the *individual* variation in change because the individual test-takers may have different trajectories (i.e., individual differences in intra-individual change): For example, some test-takers show a decline in effort, some an increase and some either a decline or an increase in effort. The next subsections describe the research questions in more detail and point out to the type of change we focus on.

*Change in Test-Taking Motivation*

(1) Does TTM change on average within a two-hour cognitive low-stakes testing session?

The first question pertains to the average change (a) in expectancy for success, perceived value of the test, and test-taking effort. Based on previous research that found a slight decrease in effort within one cognitive test and a stable level of perceived importance of the test (Cao & Stokes, 2008; Horst, 2010), we expect at least a slight decrease in effort and essentially no change in perceived importance (i.e., the attainment value of EVT should be stable). Although Barry and Finney (in press) found an increase in effort during the testing session, we still expect to see a decrease in effort. This is because similar to Horst (2010) we explore change in effort within a single cognitive test rather than an entire testing session including cognitive and noncognitive tests.

Previous research did not even consider the change in expectancy for success within a testing session. We assume that there are two possibilities regarding the change in expectancy. The first possibility supposes that, on average, students' expectancy for

success should not change much across the three time points. Based on EVT, the average level of expectancy for success should remain stable because students who know the domain in which they are being tested can also estimate their corresponding competence level. Thus, students should not change much in expectancy throughout the testing session (little intraindividual change). As such, the average expectancy across test-takers at the three time points should be approximately the same (a). The second possibility assumes the presence of change in expectancy for success within a test. According to the demands-capacity model of test-taking effort (Wise & Smith, 2011), the expectancy for success could change due to the completion of previous test items. Given individual differences in students' expectancies, there could also be interindividual differences in change across test-takers (some increasing in expectancy, some decreasing in expectancy, some staying the same, as described in b). However, when averaged over test-takers (a), such differences manifest as three averages that are very similar in magnitude (because some test-takers increase, whereas others decrease and yet others stay the same). Due to the two theoretical possibilities described above, we cannot form a hypothesis about the change in expectancy.

(2) Is the change in some TTM constructs related to the change in other TTM constructs?

This research question focuses on the relationship between individual differences in change in one variable and individual differences in change in another variable (b). Previous research has found no relationship between the change in effort and the change in value, or between expectancy and change in effort in a testing session with different test types (Barry & Finney, in press). Moreover, previous research has shown that effort and value were related to each other on a cognitive test (Barry & Finney, in press). Because this study focused on the change of TTM within one cognitive test with a constant level of mental taxation (in contrast to a testing session including also noncognitive tests) we assume that the change in effort and the change in value are related. Although we expect a stable level of value on average (a; see research question 1), it is possible that students show different trajectories, similar to the second hypothesis for the change in expectancy (b). Based on the demands-capacity model of test-taking effort, that relates the change in confidence to solve future items (i.e., the change in expectancy) with the change in effort we assume that the change in expectancy for success and the change in effort are related. To understand the mechanism driving these relationships between the TTM constructs it is important to examine not only change on average, but individual change trajectories.

*Change in TTM and Its Relationship to Test Performance*

(3) What is the relationship between change in TTM and test performance after accounting for students' socio-demographic background, ability in mathematics, and domain-specific motivation?

According to EVT, expectancy and value influence the expended effort on a given test. It is reasonable to assume that students with a decreasing level of TTM show a lower test performance relative to students with a stable level of TTM. In particular, as the level of effort declines, students may answer only easy items or abandon the test altogether, which would manifest in a low test score. In contrast, change in effort may not be related to test performance because those students who decrease in effort over time may still score higher or lower on the total test than those students who remain stable in effort over time. Thus, the relationship between change in effort and total test performance is difficult to interpret.

## 6.3 Method

*Sample*

The current study explores TTM in the German National Assessment Study conducted by the Institute for Educational Quality Improvement (Pant et al., 2013). This study is a typical low-stakes assessment, and it measured mathematical and scientific literacy in a representative sample of German ninth-graders ($N$ = 44,584). Students with special needs ($N$ = 1,380) were excluded because they received a different test from the rest of the sample. In addition, students were excluded if they intentionally disregarded the instructions of the TTM questionnaire (i.e., handing in a completely blank questionnaire or providing the same response option on all items; $N$ = 906). Thus, a total of 42,298 students were included in this study. About half the sample (49.8%) was female, and the mean age was 15.6 years ($SD$ = 0.61). One third of the students (35.2%) attended the academic-track; the remaining students were assessed at non-academic track schools. Germany has a tracked school system. After elementary school, German students are assigned to different school tracks, primarily according to their performance. The academic-track is the *Gymnasium* and leads to the degree that is prerequisite for university education (Trautwein, Lüdtke, Marsh, Köller, & Baumert, 2006). The non-academic track encompasses heterogeneous school types that are academically less demanding than the *Gymnasium*. In some school types of the non-academic track, students are prepared for apprenticeships; in other

school types, students receive a more comprehensive general education. The *Gymnasium* is the only school type that exists in all German federal states within an otherwise very heterogeneous school system (Pant et al., 2013).

*Procedure and Instruments*

*Procedure*. The assessment took place in spring 2012, and TTM was assessed at three measurement points. The TTM items were presented for the first time after the general instructions for the test, which included sample items and information about the content of the test (T1). Following the TTM questionnaire, students worked on the first half of the achievement test. After the first test half (one hour) and a 15 minute break, students completed the TTM items again (T2). After finishing the second test half and taking another break, students completed a socio-demographic background questionnaire. For approximately half of the sample, this questionnaire contained items on TTM (T3). In summary, during the two-hour testing session, TTM was assessed three times: before the test, after the first half of the test, and (for half of the sample) after the test.

*Achievement test*. The test is a standards-based assessment designed to evaluate and compare the competencies of students in the German federal states. The achievement test (*National Assessment Study* 2012; Pant et al., 2013) assessed mathematical and scientific competencies, but our study focused only on mathematics. The mathematics test consisted of 374 items (39% multiple choice, 11% constructed short-response, and 50% open-ended). A balanced incomplete block design was used (i.e., every student was administered only a subset of all items; Frey, Hartig, & Rupp, 2009), in order to administer a sufficient number of items of the test domain within a limited testing period. Weighted likelihood estimates (Warm, 1989) were computed as measures of test performance in mathematics (our last research question). These estimates are based on unidimensional scaling with the 1-parameter logistic (Rasch) model. The test was a typical low-stakes assessment for the test-takers because they did not receive a grade or individual feedback on their test performance. The mathematics test showed a high reliability (WLE Person separation reliability = .91; EAP/PV reliability = .91).

*Test-taking motivation*. We used the Questionnaire on Current Motivation to measure both the expectancy and value components of EVT (Freund, Kuhn, & Holling, 2011; Rheinberg, Vollmeyer, & Burns, 2001). The scale of the items ranged from 1 = strongly disagree to 4 = strongly agree. The expectancy component is represented by the perceived

probability of success (e.g., “I think I am up to the difficulty of this test”), and the value component is represented by the challenge subscale (e.g., “I am eager to see how I will perform on the test”). The concept of challenge denotes the extent to which test-takers perceive the situation as an achievement situation and corresponds to the attainment value in EVT (i.e., perceived importance of the test). The challenge subscale is henceforth referred to as the importance subscale. Probability of success refers to the students’ confidence in doing well on the test (Freund et al., 2011; Rheinberg et al., 2001). At the beginning, the questionnaire assesses the current motivation before taking the cognitive test. Then, the items were adapted to assess TTM after the first half of the test and again after the test. Additionally, test-taking effort was assessed with three items from the TTM scale by Eklöf (2010b; e.g., “I worked on each item in the test and persisted even when the task seemed difficult”). The scale of these items also ranged from 1 = strongly disagree to 4 = strongly agree. All scales relate to the current test situation and aim to measure states (as opposed to traits). The effort scale was originally developed to measure invested effort after a test, but these items were also adapted for the first and second time point.

The measurement model for all five TTM constructs was established by Penk and Schipolowski (2014). Confirmatory factor analyses of the first and third measurement point demonstrated a very high correlation between the effort and importance factors at each time point. This strong correlation was mostly likely due to the fact that one item from the importance scale (“I am really going to try as hard as I can on this test”) measures effort. As a consequence, this item was included in the effort factor instead in the importance factor. Thus, our study used a four-item effort scale, a three-item expectancy scale, and a two-item importance scale. The descriptive statistics of the TTM subscales are displayed in Table 6.1. The values of the three subscales indicated a decrease in effort and importance during the test. In contrast, probability of success appeared to be stable. We estimated reliability using McDonalds’s Omega (McDonald, 1999), which can be interpreted similarly to Cronbach’s Alpha. This measure is appropriate, because it considers the standardized factor loadings of the items from the confirmatory factor analysis and thus does not assume equal factor loadings as Cronbach’s Alpha does (Sijtsma, 2009). The reliability estimates for the factors were acceptable given the small number of items on each scale.

Table 6.1

*Descriptive Statistics and the Reliability of the TTM Scales*

| Scale | | T1 | | | T2 | | | T3 | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | $N_{items}$ | *M* | *SD* | ω[1] | *M* | *SD* | ω[2] | *M* | *SD* | ω[3] |
| Effort | 4 | 2.95 | 0.67 | .83 | 2.73 | 0.75 | .86 | 2.55 | 0.76 | .85 |
| Probability of success (E) | 3 | 2.88 | 0.56 | .64 | 2.85 | 0.66 | .70 | 2.85 | 0.64 | .67 |
| Importance (V) | 2 | 2.75 | 0.80 | .67 | 2.47 | 0.86 | .73 | 2.39 | 0.87 | .75 |

Notes: *M* = mean; *SD* = standard deviation; ω = McDonald's Omega; E = expectancy for success; V = value; T1–T3 = measurement times before the test, half way through the test, and after the test, respectively. [1]$N_{T1}$ = 42,080; [2]$N_{T2}$ = 42,099; [3]$N_{T3}$ = 22,601.

*Student Background Questionnaire*. Students completed an instrument with self-report scales at the end of the test. This questionnaire included questions on the student's home environment, self-related beliefs, and instructional quality in the classroom. In this study, we used the data for students' self-concepts in mathematics as a measure of domain-specific motivation. This construct was measured with four items (e.g., "I get good grades in mathematics"; Ramm et al., 2006) on a four-point Likert-scale ranging from 1 = strongly disagree to 4 = strongly agree (McDonald's ω = .91). Five variables from the student background questionnaire were used as control variables: (a) sex, (b) school track, (c) socio-economic status, (d) immigration background, and (e) grade in mathematics. Previous research has found differences in TTM between academic-track students and nonacademic-track students, so we included the school track as a control variable in our analyses (Penk, Pöhlmann, & Roppelt, 2014). Socio-economic status was assessed with the Highest International Socio-Economic Index of Occupational Status (HISEI; Ganzeboom, De Graaf, & Treiman, 1992), which is an indicator of the status of the parents' professions with respect to income and education level. The HISEI scores were standardized. Students' immigration background was defined according to Stanat and Chistensen (2006). Specifically, student has an immigration background if (a) one parent was not born in Germany, (b) both parents were not born in Germany, but the student was born in Germany, or (c) both parents and the student were not born in Germany. Students' grade in mathematics (standardized and centered at their class mean) was included to control for students' relative ability in mathematics before taking the test. Due to Germany's grading system lower values indicate a higher ability level than higher values.

*Analyses*

In this study, second-order latent growth curve modeling was used to examine the trajectories of the TTM constructs within a test. This modeling framework allows for estimating the initial level of TTM as well as the change in TTM at three measurement points. It makes it possible to examine intraindividual change in expectancy, value, and effort during the test, as well as interindividual differences in this intraindividual change (Sayer & Cumsille, 2001). *Second-order* latent growth curve models use latent variables to estimate growth over time (e.g., a latent motivation variable composed of four effort items measured at three time points). Thus, the latent variables form the *first-order factors*, and the growth parameters form the *second-order factors* (Ferrer, Balluerka, & Widaman, 2008; Sayer & Cumsille, 2001). See Figure 6.2. This technique allows for the separation of measurement error from true trait change and reliable time-specific variance. In addition, these models also make it possible to test the assumption of measurement invariance over time. Finally, they have the statistical power needed to uncover individual differences in change (Geiser, Keller, & Lockhart, 2013; Sayer & Cumsille, 2001).

Knowing the advantages of second-order latent growth curve modeling we describe now the model specification. Latent growth curve modeling includes an intercept and one or more slope factors as growth factors. Due to the three measurement points in this study, we could only estimate linear trajectories (i.e., one linear slope factor) and could not test more complex shapes of growth, such as quadratic or piecewise growth (Hancock, Kuo, & Lawrence, 2001). The intercept factor represents the initial level of the variable of interest (e.g., effort before the test). As shown in Figure 6.2, the paths from the intercepts (e.g., initial effort) to the three latent first-order factors (e.g., effort variables at three time points) are fixed at 1, because the intercept is a stable constant without growth. The slope factors describe the linear rate at which the variable of interest changes over time, e.g., a decrease in effort during the testing session (Preacher, Wichmann, MacCallum, & Briggs, 2008). Hence, the paths from the slope factor (e.g., change in effort) to the three latent first-order factors (e.g., effort variables at three time points) are fixed at 0, 1, and 2, respectively, reflecting the linear trajectories. The path from the slope factor to the first time point is fixed to zero because there can be no growth at the initial time point.

*Change in test-taking motivation*. Three linear growth models are estimated to answer the first research question investigating change in TTM during the testing session: one model each for probability of success, perceived importance, and test-taking effort. All

covariances among the first-order factors (e.g., covariances among the latent effort variables at the three time points) were set to zero under the assumption that the relations among the first-order factors are explained fully by the second-order latent growth factors (Sayer & Cumsille, 2001). For the second research question examining whether change in the three TTM constructs is related to each other, we simultaneously estimated the growth processes of probability of success, importance, and effort using one multivariate second-order latent growth model. This model allows correlations among all of the growth parameters of the three TTM constructs. Thus, all of the intercept and slope factors for probability of success, importance, and effort were estimated in one model and were allowed to correlate.

*Change in TTM and its relationship to test performance*. For the last research question, we used a two-step procedure estimating two consecutive models. We first examined how much variance in the test scores could be explained by students' socio-demographic background (i.e., sex, school track, socio-economic status, and immigration background) variables (Model 1). In the next step, we predicted test performance with the growth parameters of TTM within the testing session. Thus, we added the intercept and slope factors for expectancy, importance, and effort as predictors of test performance in mathematics. The self-concept in mathematics was added as a predictor of both test performance and probability of success to control for any spurious relationship between the two constructs due to both being related to self-concept (Eccles & Wigfield, 2002; Eklöf, 2006, 2007, 2008). The intercepts and slopes for perceived importance and probability of success were used as predictors of the corresponding growth factor of effort. In this way, we want to test primarily whether the effect of change in importance and change in probability of success on test performance is mediated by change in effort (Cole et al., 2008; Penk & Schipolowski, 2014). The final model explores how much of the variance in test performance is associated with the state-like TTM constructs, while controlling for the students' socio-economic background characteristics and domain-specific motivation.

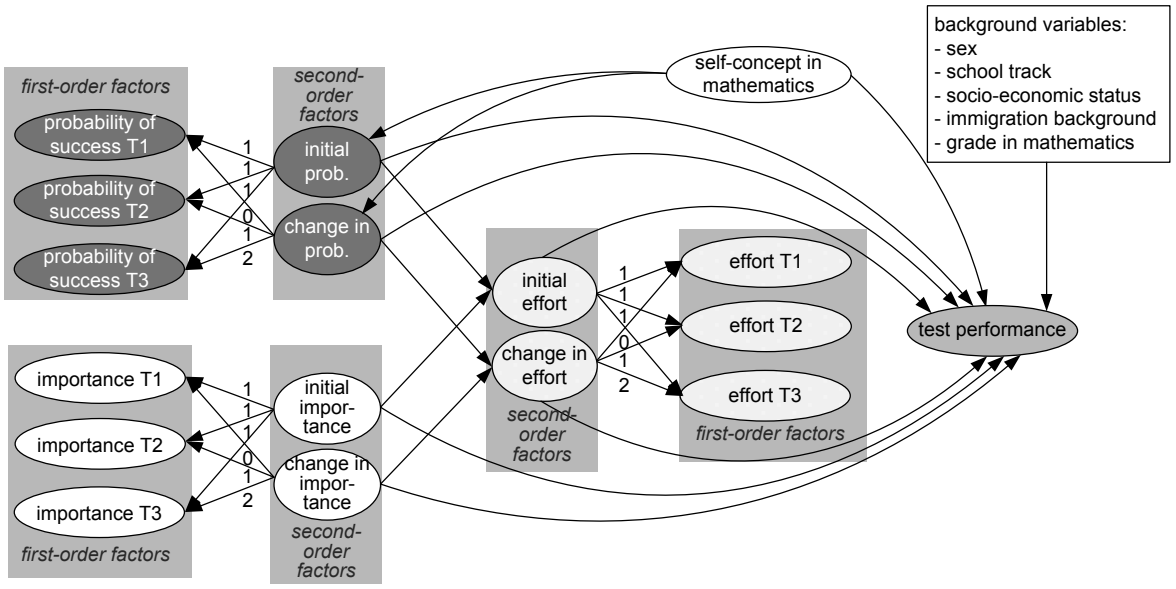



*Figure 6.2*. Prediction of test performance with the growth factors for probability of success, importance, and effort, controlling for students' background characteristics and self-concept in mathematics. Notes: Manifest indicators, disturbance terms, and correlations are omitted for simplicity. prob. = probability of success; T1–T3 = measurement times before the test, half way through the test, and after the test, respectively.

Analyses were conducted in M*plus* 7.1; (Muthén & Muthén, 2012). The hierarchical structure of the data (i.e., students nested within classes) was taken into account in the computation of standard errors and model fit. In addition, sample weights were used in all of the analyses to ensure the results are representative of the population of ninth-graders in German schools. Due to the large sample size (i.e., high power), we specified a *p*-value below .001 as the cut-off for statistical significance. In all of the analyses, we applied a robust maximum likelihood estimator and considered the following indices to evaluate the model fit: (a) MLR $\chi^2$-statistic, corresponding degrees of freedom, and probability value; (b) comparative fit index (CFI); (c) Tucker-Lewis index (TLI); (d) root mean square error of approximation (RMSEA); and (e) standardized root mean square residual (SRMR). According to Hu and Bentler (1999), the following values indicate adequate model fit: CFI > .95, TLI < .95, RMSEA < .06, and SRMR < .08. Strong measurement invariance is required to apply second-order latent growth curve modeling. This requirement ensures that the items in the questionnaire assess the same construct at every measurement point. To compare the different measurement invariance models (configural, metric, and strong measurement invariance), we used ΔCFI (Cheung & Rensvold, 2002; Rutkowski & Svetina, 2014) and the Root Deterioration per Restriction statistic (RDR; Browne & Toit,

1992), comparing the relative fit of nested models based on their RMSEA differences. Values of ΔCFI < -.01 and values of RDR < .05 suggested a good model fit. As mentioned by Barry and Finney (in press), although we apply these general guidelines to evaluate the structural equation models, one should keep in mind that there is currently little research on applying these indices to second-order latent growth models.

## 6.4 Results

Before presenting the results, it is important to test whether the data meet the requirements for the application of second-order latent growth modeling. The model requires that the assumption of strong measurement invariance holds. More specifically, the constructs of interest need to be represented by the same structure over time (configural measurement invariance), the same factor loadings over time (metric measurement invariance), and the same item intercepts over time (strong measurement invariance). Appendix A includes the results of these nested measurement invariance models. The constructs used in this study exhibited strong measurement invariance (effort: $\chi^2(51) = 743.16$, $p < .001$; CFI = .99; TLI = .99; RMSEA = .02; SRMR = .03; RDR = .03; probability of success: $\chi^2(23) = 265.89$, $p < .001$; CFI = .99; TLI = .99; RMSEA = .02; SRMR = .03; RDR = .03). Strict measurement invariance (in addition to strong measurement invariance, the variables have the same error variances over time) is generally unlikely in growth models due to heterogeneous variance over time (Sayer & Cumsille, 2001). The importance scale consists of only two items. As invariance tests require at least three indicators, we cannot use goodness-of-fit statistics to evaluate model fit (Bollen, 1989). However, the second-order latent growth model assuming strong measurement invariance for the importance construct fits the data well ($\chi^2(7) = 547.12$, $p < .001$; CFI = .98; TLI = .97; RMSEA = .04; SRMR = .03). This suggests that the importance factor also exhibits strong measurement invariance.

*Change in Test-Taking Motivation*

The first research question addressed change in TTM over a two-hour low-stakes testing session. As shown in Table 6.2, all three second-order latent growth curve models estimating the linear change in effort, importance, and probability of success showed satisfactory fit. The mean of the effort intercept was 2.98 on a 4-point scale, indicating that, on average, students reported that they were willing to invest effort before the test. The coefficients in Table 6.2 differ slightly compared to the coefficients in Table 6.1 because

the former are latent values and the latter manifest values. The mean of the linear slope was -0.13 (β = -.81), indicating that students' average effort decreased during the testing session as hypothesized. After half of the test, the mean average effort decreased to 2.84 [≈ 2.98 + (1 × -0.13)], and at the end of the test, the mean average effort was 2.71 [≈ 2.98 + (2 × -0.13)]. The variances of the intercept and slope factors were statistically significant, but students showed much more variability in their initial effort than in their change in effort. Assuming that the growth parameters followed a normal distribution, the estimated means and variances can be used to generate a distribution of the change in effort. Approximately two-thirds of the students showed slope values between -0.30 and 0.03; that is, some showed a greater decrease in effort, whereas others remained more or less stable. The correlation between the intercept and slope for effort was negative and statistically significant ($r$ = -.14), indicating that students with a higher initial effort than average had a greater decrease in effort than average. However, the correlation was quite small.

The growth parameter estimates for the importance factor were similar to the estimates for effort. The mean of the importance intercept was 2.70, indicating that, on average, students perceived the test as important before they took the test. The mean of the importance slope was negative and statistically significantly -0.18 (β = -.73). In other words, contrary to our hypothesis the level of importance decreased significantly during the testing session. Again, the variances of the intercept and slope factors were significant, and students showed greater variability in their initial importance than in change in importance. The variability in importance, especially before the test, was very high, indicating students valued the test quite differently from one another. However, about two-thirds of the students demonstrated slope factors ranging from -0.43 to 0.07, which indicated that some students perceived the test as important during the entire test, whereas for others importance decreased more than the average decrease. In contrast to the effort growth parameters, the importance intercept and slope were not significantly correlated.

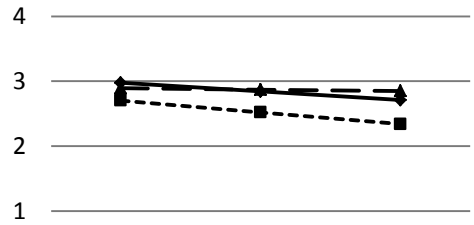
The initial perceived probability of success was 2.89, which is similar to the initial levels of the other two constructs. Prior to the test, the average student felt confident that they would complete the test successfully. However, the mean of the slope of probability of success was almost 0 (β = -.11), though statistically significant, indicating that students' probability of success remained mostly stable. The variances of the intercept and slope factors were significantly different from zero, and again students showed more variability

in the initial probability of success than in change in the probability of success. Considering the standard deviation, approximately two-thirds of the students had a mean slope ranging between -0.21 and 0.16, indicating that some students reported a decrease in their perceived probability of success, whereas others reported an increase. The intercept and slope for probability of success were not correlated.

Table 6.2

*Parameter Estimates and Model Fit for the Second-Order Latent Growth Curve Models for Effort, Importance, and Probability of Success*

| Factor | | *M* | *SD* | Correlation between intercept & slope | T1 | T2 | T3 |
|---|---|---|---|---|---|---|---|
| Effort | intercept | 2.98* | 0.47* | -.14* | | | |
| | slope (solid line) | -0.13* | 0.16* | | | | |
| Impor-tance | intercept | 2.70* | 0.78* | | | | |
| | slope (dotted line) | -0.18* | 0.25* | -.06 | | | |
| Probability of Success | intercept | 2.89* | 0.41* | | | | |
| | slope (dashed line) | -0.02* | 0.18* | -.05 | | | |




| | $\chi^2$ *(df)* | CFI | TLI | RMSEA | SRMR | *N* |
|---|---|---|---|---|---|---|
| Effort | 971.64* (54) | .99 | .99 | .02 | .03 | 42,287 |
| Importance | 547.12* (7) | .98 | .97 | .04 | .03 | 42,281 |
| Probability of Success | 415.88* (25) | .99 | .98 | .02 | .04 | 42,292 |

Notes: *$p < .001$. *M* = mean; *SD* = standard deviation; T1–T3 = measurement times before the test, half way through the test, and after the test, respectively; *df* = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual.

The second research question focused on the relationship between the growth parameters for effort, importance, and probability of success. For this purpose, we modeled the three growth processes simultaneously in one multivariate second-order latent growth curve model. Table 6.3 contains the factor correlations. The estimated model fitted the data well ($\chi^2(311) = 7781.83$, $p < .001$; CFI = .96; TLI = .95; RMSEA = .02; SRMR = .06). All

of the slope factors were positively correlated with each other. The intercept of the effort factor and the intercept of the importance factor showed the highest correlation, as did the slope of the effort factor and the slope of the importance factor ($r$ = .79 for both). As both slopes were negative, the correlation expresses that a smaller decrease in importance for students is associated with a smaller decrease in test-taking effort over the testing session. In other words, students who decrease more in effort relative to other students tend to decrease more in importance relative to other students. The correlation between the slopes for probability of success and effort was also significant ($r$ = .45) and indicated that a smaller decrease in probability for success tends to be accompanied by a smaller decrease in test-taking effort. Additionally, the slopes for importance and probability of success were moderately correlated ($r$ = .33), indicating that the smaller the decrease in probability of success, the smaller the decrease in importance. Moreover, the intercepts for effort and probability of success showed a small correlation ($r$ = .21), indicating that test-takers who decreased more than average on probability of success tended to decrease more than average on effort.

Table 6.3

*Correlations Between the Intercept and Slope Factors of the Multivariate Second-Order Latent Growth Curve Model with Effort, Importance, and Probability of Success*

| Factor | Factor correlations | | | | | |
|---|---|---|---|---|---|---|
| | 1. | 2. | 3. | 4. | 5. | 6. |
| 1. Effort intercept | – | | | | | |
| 2. Effort slope | -.12* | – | | | | |
| 3. Importance intercept | .79* | -.04 | – | | | |
| 4. Importance slope | -.08* | .79* | -.04 | – | | |
| 5. Probability of success intercept | .21* | -.04 | .00 | -.01 | – | |
| 6. Probability of success slope | .08* | .45* | .09* | .33* | .01 | – |

Notes: *$p$ < .001.

To sum up the first two research questions, the three second-order latent growth curve models showed a moderate initial TTM before the test and a moderate decrease in effort and importance within the test. In contrast, students' probability of success remained stable. The slopes of the three TTM constructs were significantly correlated with each

other, indicating moderate to strong relationships between the changes in effort, importance, and probability of success.

*Change in TTM and Its Relationship to Test Performance*

The last research question investigated the change in TTM and its relationship to test performance in a two-step procedure. First, we predicted the mathematics score solely with the student's background information: sex, school track, socio-economic status, immigration background, and grade in mathematics (Model 1). This model did not include any motivational variables. In the second step, all of the growth parameters of the three TTM constructs as well as the domain-specific motivation were added as predictors of test performance (Model 2).

The first step of the procedure is analyzed in Model 1. The model showed good fit ($\chi^2(25) = 169.96$, $p < .001$; CFI = .99; TLI = .99; RMSEA = .02; SRMR = .01). The five background variables significantly predicted the mathematics scores and explained 57% of their variance. The strongest predictor was school track ($\beta = .59$), indicating that students attending the academic track outperformed their classmates in non-academic tracks. The grade in mathematics ($\beta = -.33$) also predicted students' test performance. Sex ($\beta = .12$), immigration background ($\beta = -.12$), and socio-economic status ($\beta = .10$) also significantly predicted test performance, but these coefficients were quite small. Specifically, male students outperformed female students, students without an immigration background outperformed students with an immigration background, and the higher the socio-economic status, the higher the test score of the student.

The second model (presented in Figure 6.3) also fit the data well. The three TTM constructs, students' background variables, their grade in mathematics, and their domain-specific motivation explained 64% of the variance in mathematics scores. The latent growth curve models of the three TTM constructs and the domain-specific motivation explained an additional 7% of the test score variance.

Looking at the paths in the model in more detail, self-concept in mathematics was a predictor of test performance as well as a strong predictor of the probability of success intercept. Specifically, self-concept in mathematics explained almost a quarter of the variance in initial probability of success. Yet, self-concept did not predict the change in probability of success. After controlling for self-concept in mathematics, the probability of success intercept and slope significantly predicted test performance; the higher the initial

probability of success and the smaller the decrease in probability of success (or the greater the increase), the better the student's test performance.

The intercept and slope for importance had no significant direct effects on test performance, but they significantly predicted the respective effort factors. Over 60% of the variance in the growth factors for effort could be explained by the growth factors for importance and probability of success. However, the intercept and slope for effort were mainly predicted by the intercept and slope for importance. In addition, the intercept for effort significantly predicted test performance. Thus, the higher the initial effort, the better the student's performance.

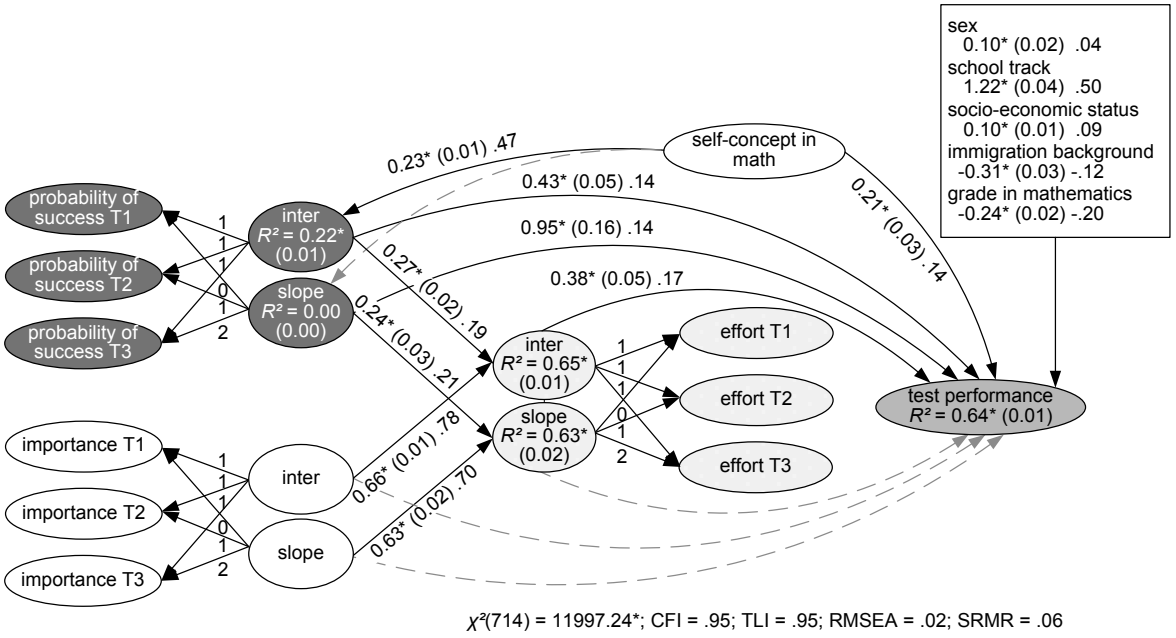



*Figure 6.3*. Prediction of test performance with the growth factors for probability of success, importance, and effort, controlling for the students' backgrounds and self-concepts in mathematics (Model 2). Notes: Manifest indicators, correlations, and the measurement and residual errors are omitted for simplicity. Coefficients: *unstandardized coefficient* with *p (standard error) standardized coefficient*. *$p < .001$. inter = intercept; T1–T3 = measurement times before the test, half way through the test, and after the test, respectively; T3 = after the test. Non-significant paths are dashed, but the coefficients are listed in Appendix B.

Of the indirect effects shown in Appendix B, only the effect of the intercept for importance on test performance via effort was substantial and significant ($\beta = .13$). Thus, the effect of initial importance on test performance was fully mediated by the initial level of

effort. The indirect effect of the initial probability of success on test performance was significant but quite small ($\beta = 0.03$). Moreover, the intercept and slope for probability of success showed significant and substantial direct effects on test performance. Neither was mediated by effort.

In summary, the probability of success intercept and slope and the effort intercept were direct predictors of the mathematics scores after accounting for students' background characteristics. The importance intercept and slope were not directly related to test performance, but they directly predicted the test-taking effort intercept and slope, respectively. Effort mediated only the effect of the importance intercept on test performance. Both the intercept and slope for probability of success predicted test performance after controlling for domain-specific motivation. Thus, there appears to be a relationship between the state variable probability of success and test performance, beyond the trait-like variable self-concept in mathematics. Additionally, the test-taking effort intercept was the strongest predictor of test performance among the parameters of the TTM constructs. The sizes and, therefore, the effects of the significant growth parameter coefficients on test performance were comparable to the background variables: socio-economic status and immigration background.

## 6.5 Discussion

As the number of low-stakes tests in German schools has increased, research on test-taking motivation (TTM) has grown in the last decade. Research shows that test scores from low-stakes assessments may be affected by low motivation (Cole et al., 2008; Eklöf & Nyroos, 2013; Swerdzewski et al., 2011; Wise & DeMars, 2005; Wolf & Smith, 1995). Understanding the mechanism of TTM during the testing session and the effects of TTM on the test scores is crucial to ensure the proper interpretation of test results (Thelk et al., 2009). Thus, the current study had two main purposes. First, we explored the change in three TTM constructs based on expectancy-value theory (EVT). Specifically, we examined probability of success (expectancy), perceived importance of the test (attainment value), and test-taking effort within a testing session in which students completed a cognitive test in mathematics. Our second goal was to investigate the relationship between change in the three TTM constructs and students' test performance.

*Change in Test-Taking Motivation*

The first research question addressed the change in probability of success, perceived importance of the test, and test-taking effort. The results showed that, on average, probability of success remained stable over the testing session, but perceived importance of the test and test-taking effort decreased within the testing session. Despite the significant variability in the change in all three TTM constructs, it can be considered "good news" that, on average, students reported a moderate decrease on two of the three TTM constructs, although they completed a two-hour cognitive (and mentally demanding) test without any personal consequences. The moderate decrease might be due to the break students had to recover from the first half of the test. It seems that a two-hour low-stakes test is an adequate time frame, considering intraindividual motivational processes. The results of our study are in line with Horst's (2010) findings, which demonstrated a slight decrease in test-taking effort on one cognitive test ($d$ = 0.19) over a 50-minute period. The students in our study reported a slightly larger change in the effort factor ($d_{T1/T2}$ = 0.31; $d_{T2/T3}$ = 0.23) than the students in Horst's study. However, the Horst study did not include a measure of TTM before the test, as we did in our study. However, on average, probability of success remained at a stable level. Thus, although students were asked before the test, it appears that they provided realistic estimates of their expectancy for success. In contrast to Horst's (2010) results, the importance scale in our study showed a decrease similar to that of effort. This may be attributable to the fact that we measured the attainment value of EVT indirectly using the challenge scale, instead of directly measuring the perceived importance of the test.

Overall, test-takers showed more variability in their initial TTM than in their change in TTM. The variability in initial importance was especially high, indicating that students varied a lot in their perceived value of this test. Although beginning at different levels of TTM, on average students showed similar change in TTM throughout the test. Thus far, no other study has investigated change in the value and expectancy components within one cognitive test, so we cannot compare the findings with those of other investigations. However, Horst's (2010) results indicated a fairly stable level for importance and lower variances for the three importance means in comparison to the effort means. The variability of change in the TTM constructs was small, but nevertheless significantly different from zero.

Before investigating the relationship between change in TTM and test performance, we explored whether the changes in probability of success, perceived importance of the test, and test-taking effort were related to each other. The results showed a strong relationship between initial importance and initial effort, as well as between the change in importance and change in effort. If students valued the test and retained this attitude throughout the test, they were more willing to invest effort. Moreover, initial probability of success was related to initial effort. Although we found that, on average, level of probability of success was stable throughout the test, change in probability of success was related to change in effort. Thus, although on average probability of success was stable, it appears that students' individual trajectories vary enough for change in probability of success to correlate with change in effort. Effort at the beginning of the test was also strongly associated with the initial perceived importance of the test. Additionally, change in effort was also highly related to change in the perceived value of the test during the testing session. In contrast to Barry and Finney (in press), who found no relationship between change in effort and change in importance during different cognitive and noncognitive tests, our study discovered evidence that changes in the TTM constructs were related to each other. Thus, *throughout a single cognitive test*, the different TTM constructs seem to be related, in contrast to TTM over different types of tests, which showed no relationship between change in one construct and change in another. Thus, from a theoretical and practical point of view it is important to assess all three components of EVT to capture the whole growth process of TTM. It appears that students show a smaller decline in effort than average if they also report a smaller decrease in perceived importance of the test throughout the testing session.

*Change in TTM and Its Relationship to Test Performance*

The last research question focused on the relationship between change in TTM and students' test performance, after controlling for students' backgrounds. Over 50% of the variance in mathematics performance was explained by students' background characteristics, with school track and grade in mathematics being the most predominant predictors. The final model added the growth parameters of probability of success, importance, and effort as well as self-concept in mathematics as predictor of test performance and the growth parameters of probability of success. In addition to reported effort before the test, the initial level of expectancy and the change in expectancy, while controlling for self-concept in mathematics, also predicted test performance. Although the average change in

probability of success over the testing session was fairly stable the interindividual variability of intraindividual change in probability of success was high enough that it revealed a relationship with test performance. Students who decrease less in their probability of success than the average tend to score higher in the cognitive test. This finding is consistent with the results of Penk and Schipolowski (2014) who showed that the expectancy component measured after the test showed a stronger relationship with test performance than the expectancy component before the test.

Moreover, it is well known that a domain-specific self-concept is related to performance in the corresponding domain (Chen, Yeh, Hwang, & Lin, 2013; Jansen, Schroeders, & Stanat, 2013) and can affect probability of success (Asseburg, 2011; Eccles & Wigfield, 2002). Our study showed that beyond the stable domain-specific motivation, it is also important that students feel confident they will complete the test successfully. This result is consistent with the demands-capacity model of test-taking effort (Wise & Smith, 2011). The completion of previous test items can change the confidence in successfully completing the test, and in turn, the amount of effort students are willing to invest in further test items (Wise & Smith, 2011). Thus, a test booklet that includes alternating easy and difficult items throughout the test might be necessary to ensure a stable level of expectancy for success as well as high test performance among students.

Our study found an effect only for change in expectancy for success on test performance. However, change in importance of the test was an important predictor of change in effort. The proctor strategies presented by Lau and colleagues (2009) could be a promising means of increasing the perceived value of the test, and in turn, students' effort. The authors found that motivation enhancing behavior of the proctors (invigilators; such as emphasizing the importance and usefulness of the test; encouraging test-takers to give their full effort during the testing session) during the low-stakes testing session can affect invested effort. Emphasizing the importance and usefulness of the test aligns with emphasizing the value component of EVT. Our study provided support that attainment value and effort are strongly related in that students who valued the test also showed higher effort and better test performance. We therefore recommend at least emphasizing the importance and usefulness of the test in the test instructions. In addition, the moderate change in TTM throughout the cognitive test indicated that a two-hour low-stakes test was not too exhausting for the students, although a longer test time might lead to a larger decrease in TTM. In this case, a further decrease in TTM could be due to the amount of fatigue or time pressure

felt by students, which in turn might affect their willingness to invest effort on further items (Wise & Smith, 2011).

*Limitations and Directions for Future Research*

There are several limitations of our study that we address below. Although the strong measurement invariance of the TTM scales supported a successful adaption of the test items to different measurement points, more items per subscale would have been preferable (especially for the importance factor). Four items per subscale seems to be appropriate for second-order latent-growth modeling. Moreover, more measurement points are desirable, in order to test different growth forms in addition to linear change in TTM. For example, it is possible that a piecewise growth form fits the data in our study better, such as a larger decrease in TTM during the first half of the test and a smaller decrease in TTM during the second half of the test.

Furthermore, the measurement of the value scale can be optimized. We assessed indirectly the attainment value using the challenge scale. Most of the studies conducted internationally (Cole et al., 2008; Eklöf & Nyroos, 2013; Eklöf et al., 2013; Thelk et al., 2009; Wolf & Smith, 1995) assess the attainment value directly by asking students about their perceived importance of the test. It is conceivable that asking students directly how they perceive the test would lead to somewhat different results. Moreover, and as described above, the value component consists of four different aspects. In this study, only one aspect was included in the analyses. This is an opportunity for future research. Furthermore, this study used self-report measures of TTM. As stated by Swerdzewski and colleagues (2011), such measures have several disadvantages: the test-takers a) need to recognize their current level of TTM, b) need to use the scale accurately to express their TTM, and (c) need to truthfully report their TTM. Despite these limitations, self-report measures are quite common in large-scale paper-and-pencil assessments.

Another limitation concerns the consideration of students' previous ability. Most previous research has found that the level of effort is not substantially related to high-stakes test scores when cognitive ability is controlled (DeMars, Bashkov, & Socha, 2013; Kong, Wise, Harmes, & Yang, 2006; Wise, Bhola, & Yang, 2006; Wise & Kong, 2005), but moderately related to low-stakes test scores (DeMars et al., 2013). Our study used students' grade in mathematics as a measure of students' ability in mathematics prior to the test. We know that school grades account for not only intellectual capacity, but also for

motivational and personality aspects; thus, grades are less objective than test scores on standardized achievement tests (Spinath, 2012). Ideally, we would like to control for prior knowledge with an additional measure from a high stakes test; however, this information was not available to us.

Furthermore, we did not assess TTM for the completion of the student questionnaire like previous studies (Barry & Finney, in press; Barry et al., 2010; Horst, 2010). Instead, we used students' answers to draw conclusions about their attitude about school or to determine their socio-economic status. Thus, it is important that students also complete these questions with high effort. Further studies could compare TTM in large-scale assessments for both the cognitive test and the student questionnaire, as well as investigate change in TTM during the entire testing session, including the noncognitive test.

*Conclusions*

Our investigation of the change in TTM over the course of a cognitive large-scale assessment and its relationship to test performance based on EVT adds to the existing body of TTM research. We found an effect of TTM on test performance after taking into account students' socio-demographic background and their domain-specific motivation. Above all, it seems crucial that students begin the test with a high level of TTM and remain confident that they can complete the test successfully through the end of the testing session. To understand the mechanism of TTM during a testing session, it is important to assess all three components of EVT or one risks missing an essential TTM construct in low-stakes assessments.

## Notes

[1] Due to the non-significant, slightly negative residual variances of some first-order factors in the second-order latent growth models, we had to fix some of the residual variances of the first-order factors to zero: for effort and importance for the first and third time point, and for probability of success for the third time point. An investigation of the residual variances using latent growth modeling with a composite of the manifest indicators per time point (instead of a latent variable) showed that these residual variances were close to zero. This supported our decision to fix the corresponding residual variances to zero.

[2] β refers to the stdyx standardization in the M*plus* output using full standardization with respect to both latent and observed variables.

## References

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Asseburg, R. (2011). *Motivation zur Testbearbeitung in adaptiven und nicht-adaptiven Leistungstests [Test-taking motivation in adaptive and sequential achievement testing]* (Doctoral dissertation). Christian-Albrechts-Universität zu Kiel. Retrieved from the website http://d-nb.info/1013153863/34

Barry, C. L., & Finney, S. J. (in press). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education*.

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*(4), 342–363. doi:10.1080/15305058.2010.508569

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441–462.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Browne, M. W., & Toit, S. H. C. D. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research*, *27*(2), 269–300. doi:10.1207/s15327906mbr2702_13

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*(2), 209–230. doi:10.1007/s11336-007-9045-9

Chen, S.-K., Yeh, Y.-C., Hwang, F.-M., & Lin, S. S. J. (2013). The relationship between academic self-concept and achievement: A multicohort–multioccasion study. *Learning and Individual Differences*, *23*, 172–178. doi:10.1016/j.lindif.2012.07.021

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(2), 233–255. doi:10.1207/S15328007SEM0902_5

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*(4), 609–624.

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, *8*, 69–82.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. doi:10.1146/annurev.psych.53.100901.135153

Eklöf, H. (2006). Development and validation of scores from an instrument measuring student test-taking motivation. *Educational and Psychological Measurement*, *66*(4), 643–656. doi:10.1177/0013164405278574

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, *7*(3), 311–326.

Eklöf, H. (2008). Test-taking motivation on low-stakes tests: A Swedish TIMSS 2003 example. In *Issues and methodologies in large-scale assessments, IERI Monograph Series* (Vol. 1, pp. 9–21). Hamburg: IEA-ETS Research Institute.

Eklöf, H. (2010a). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345–356. doi:10.1080/0969594X.2010.516569

Eklöf, H. (2010b). *Student motivation and effort in the Swedish TIMSS Advanced field study*. Paper presented at the 4th IEA International Research Conference, Gothenburg.

Eklöf, H., & Nyroos, M. (2013). Pupil perceptions of national tests in science: Perceived importance, invested effort, and test anxiety. *European Journal of Psychology of Education*, *28*(2), 497–510. doi:10.1007/s10212-012-0125-6

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2013). A cross-national comparison of reported effort and mathematics performance in TIMSS Advanced. *Applied Measurement in Education*, 131127082739006. doi:10.1080/08957347.2013.853070

Ferrer, E., Balluerka, N., & Widaman, K. F. (2008). Factorial invariance and the specification of second-order latent growth models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *4*(1), 22–36. doi:10.1027/1614-2241.4.1.22

Freund, P. A., & Holling, H. (2011). Who wants to take an intelligence test? Personality and achievement motivation in the context of ability testing. *Personality and Individual Differences*, *50*(5), 723–728. doi:10.1016/j.paid.2010.12.025

Freund, P. A., Kuhn, J. T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences*, *51*(5), 629–634. doi:10.1016/j.paid.2011.05.033

Frey, A., Hartig, J., & Rupp, A. A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement: Issues and Practice*, *28*(3), 39–53. doi:10.1111/j.1745-3992.2009.00154.x

Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, *21*(1), 1–56. doi:10.1016/0049-089X(92)90017-B

Geiser, C., Keller, B. T., & Lockhart, G. (2013). First- versus second-order latent growth curve models: Some insights from latent state-trait theory. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 479–503. doi:10.1080/10705511.2013.797832

Hancock, G. R., Kuo, W.-L., & Lawrence, F. R. (2001). An illustration of second-order latent growth models. *Structural Equation Modeling: A Multidisciplinary Journal*, *8*(3), 470–489. doi:10.1207/S15328007SEM0803_7

Horst, S. J. (2010). *A mixture-modeling approach to exploring test-taking motivation in large-scale low-stakes contexts* (Unpublished doctoral dissertation). James Madison University, Harrisonburg.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing*, *1*(2), 93–114. doi:10.1207/S15327574IJT0102_1

Jansen, M., Schroeders, U., & Stanat, P. (2013). Motivationale Schülermerkmale in Mathematik und den Naturwissenschaften [Motivational characteristic of students in mathematics and science]. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Eds.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I [The IQB National Assessment Study 2012. Competencies in mathematics and the sciences at the end of secondary level I]* (pp. 347–365). Münster: Waxmann.

Kong, X. J., Wise, S. L., Harmes, J. C., & Yang, S. (2006). *Motivational effects of praise in response-time-based feedback: A follow-up study of the effort-monitoring CBT*. Presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education*, *58*(3), 196–217. doi:10.1353/jge.0.0045

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, N.J: L. Erlbaum Associates.

Moneta, G. B., & Csikszentmihalyi, M. (1996). The effect of perceived challenges and skills on the quality of subjective experience. *Journal of Personality, 64*(2), 275–310. doi:10.1111/j.1467-6494.1996.tb00512.x

Muthén, L. K., & Muthén, B. O. (1998). *Mplus user's guide. Seventh edition*. Los Angeles, CA: Muthén & Muthén.

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (2013). *The IQB National Assessment Study 2012. Competencies in mathematics and the sciences at the end of secondary level I. Summary*. Münster: Waxmann. Retrieved from http://www.iqb.hu-berlin.de/laendervergleich/laendervergleich/lv2012/Bericht/IQB_NationalAsse.pdf

Pekrun, R., Elliot, A. J., & Maier, M. A. (2009). Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance. *Journal of Educational Psychology*, *101*(1), 115–135. doi:10.1037/a0013383

Penk, C., Pöhlmann, C., & Roppelt, A. (2014). The role of test-taking motivation for students' performance in low-stakes assessments: an investigation of school-track-specific differences. *Large-Scale Assessments in Education*, *2*(1). doi:10.1186/s40536-014-0005-4

Penk, C., & Schipolowski, S. (2014). *Investigating the multiple components of test-taking motivation in a large-scale assessment context: The importance of expectancy for success*. Presented at the annual meeting of the American Educational Research Association, Philadelphia.

Preacher, K. J., Wichmann, A. L., MacCallum, R. C., & Briggs, N. E. (2008). *Latent growth curve modeling*. Los Angeles: SAGE.

Ramm, G., Prenzel, M., Baumert, J., Blum, W., Lehmann, R., Leutner, D., … Schiefele, U. (2006). *PISA 2003: Dokumentation der Erhebungsinstrumente [Documentation of the assessment instruments]*. Münster: Waxmann.

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern-und Leistungssituationen [A questionnaire for the measurement of current achievement motivation in learning and achievement situations]. *Diagnostica*, *47*(2), 57–66.

Rutkowski, L., & Svetina, D. (2014). Assessing the hypothesis of measurement invariance in the context of large-scale international surveys. *Educational and Psychological Measurement*, *74*(1), 31–57. doi:10.1177/0013164413498257

Sayer, A. G., & Cumsille, P. E. (2001). Second-order latent growth models. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (1st ed., pp. 179–200). Washington, DC: American Psychological Association.

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). *Motivation in education: Theory, research, and applications* (3rd ed.). Upper Saddle River, NJ: Pearson Education.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. doi:10.1007/s11336-008-9101-0

Spinath, B. (2012). Academic achievement. In V. S. Ramachandran (Ed.), *Encyclopedia of human behavior* (pp. 1–8). London; Burlington, MA: Elsevier/Academic Press.

Stanat, P., & Christensen, G. (2006). *Where immigrant students succeed: A comparative review of performance and engagement in PISA 2003*. Paris: Organisation for Economic Co-operation and Development.

Stanat, P., & Lüdtke, O. (2013). International large-scale assessment studies of student achievement. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 481–483). New York, NY: Routledge.

Sundre, D. L. (2007). *The Student Opinion Scale: A measure of examinee motivation: Test manual*. Retrieved from the Center for Assessment and Research Studies website: http://www.jmu.edu/assessment/resources/resource_files/sos_manual.pdf

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*(2), 162–188. doi:10.1080/08957347.2011.555217

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. *The Journal of General Education*, *58*(3), 129–151. doi:10.1353/jge.0.0047

Trautwein, U., Lüdtke, O., Marsh, H. W., Köller, O., & Baumert, J. (2006). Tracking, grading, and student motivation: Using group composition and status to predict self-concept and interest in ninth-grade mathematics. *Journal of Educational Psychology*, *98*(4), 788–806. doi:10.1037/0022-0663.98.4.788

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427–450.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, *6*(1), 49–78. doi:10.1007/BF02209024

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. doi:10.1006/ceps.1999.1015

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114. doi:10.1207/s15324818ame1902_2

Wise, S. L., Bhola, D. S., & Yang, S. (2006). *Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT*. Presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. doi:10.1207/s15326977ea1001_1

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. doi:10.1207/s15324818ame1802_2

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185–205. doi:10.1080/08957340902754650

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: science and practice in K-12 settings* (1st ed., pp. 139–153). Washington, DC: American Psychological Association.

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education, 8*(3), 227–242. doi:10.1207/s15324818ame0803_3

*Appendix A*

Table A1

*Test of the Strong Measurement Invariance Test of Test-Taking Effort and Probability of Success, with Autocorrelated Errors*

| | | $\chi^2$ (*df*) | CFI | TLI | RMSEA | SRMR | RDR | *ΔCFI* |
|---|---|---|---|---|---|---|---|---|
| Effort | Configural | 419.9* (39) | .996 | .993 | .015 | .016 | - | - |
| | Metric | 563.1* (45) | .995 | .992 | .016 | .023 | .026 | -.001 |
| | Strong | 743.2* (51) | .993 | .992 | .018 | .025 | .028 | -.002 |
| Expectancy for success | Configural | 102.2* (15) | .997 | .993 | .012 | .012 | - | - |
| | Metric | 156.9* (19) | .995 | .991 | .013 | .019 | .018 | -.002 |
| | Strong | 265.9* (23) | .991 | .986 | .016 | .025 | .027 | -.004 |

Notes: * $p < .001$. *df* = degrees of freedom; CFI = comparative fit index; TLI = Tucker-Lewis index; RMSEA = root mean square error of approximation; SRMR = standardized root mean square residual; RDR = root deterioration per restriction statistic.

*Appendix B*

Table B1

*Correlations of the Growth Parameters for Effort, Importance, Probability of Success, and Self-Concept in Mathematics: The Indirect Effects and Non-Significant Effects for Model 2*

| *Correlations* | β | *(SE)* | |
|---|---|---|---|
| Effort intercept with effort slope | -.23* | (0.00) | |
| Probability of success intercept with probability of success slope | .03 | (0.00) | |
| Importance intercept with importance slope | -.08* | (0.00) | |
| Importance intercept with probability of success intercept | -.02 | (0.00) | |
| Importance slope with probability of success slope | .34* | (0.00) | |
| Effort intercept with self-concept in mathematics | .05 | (0.01) | |
| Effort slope with self-concept in mathematics | -.03 | (0.00) | |
| Importance intercept with self-concept in mathematics | .09* | (0.01) | |
| Importance slope with self-concept in mathematics | -.01 | (0.00) | |
| *Indirect effects* | *b* | *(SE)* | β |
| Performance on importance intercept via effort intercept | 0.25* | (0.03) | .13 |
| Performance on importance slope via effort slope | 0.10 | (0.13) | .02 |
| Performance on probability of success intercept via effort intercept | 0.10* | (0.01) | .03 |
| Performance on probability of success slope via effort slope | 0.04 | (0.05) | .01 |
| *Non-significant effects* | *b* | *(SE)* | β |
| Performance on importance intercept | -0.09 | (0.04) | -.05 |
| Performance on importance slope | 0.06 | (0.18) | .01 |
| Performance on effort slope | 0.16 | (0.20) | .03 |
| Probability of success slope on self-concept in mathematics | 0.00 | (0.01) | -.02 |

Notes: *$p < .001$. *b* = unstandardized regression coefficient; *SE* = standard error;
β = standardized regression coefficient.

Table B2

*Correlations of the Growth Parameters for Effort, Importance, Probability of Success, and Self-Concept in Mathematics with the Background Variables for Model 2*

| | Sex | | School track | | Migration background | | Socio-economic status | | Grade in mathematics | |
|---|---|---|---|---|---|---|---|---|---|---|
| | β | *(SE)* | β | *(SE)* | β | *(SE)* | β | *(SE)* | β | *(SE)* |
| Effort intercept | .00 | (0.01) | .14 * | (0.02) | -.08 * | (0.02) | .10 * | (0.01) | -0.05* | (0.01) |
| Effort slope | .02 | (0.02) | .00 | (0.02) | -.09 * | (0.02) | -.01 | (0.02) | 0.01 | (0.02) |
| Importance intercept | -.14 * | (0.01) | .00 | (0.02) | .03 | (0.01) | -.04 | (0.01) | -0.08* | (0.01) |
| Importance slope | -.07 * | (0.01) | .09 * | (0.01) | -.03 | (0.02) | .06 * | (0.01) | -0.04 | (0.01) |
| Probability of success intercept | .16 * | (0.01) | .16 * | (0.02) | -.11 * | (0.01) | .11 * | (0.01) | 0.05* | (0.01) |
| Probability of success slope | -.05 | (0.02) | .34 * | (0.02) | -.06 * | (0.02) | .16 * | (0.02) | 0.04 | (0.02) |
| Self-concept in mathematics | .26 * | (0.01) | .05 * | (0.01) | -.02 | (0.01) | .06 * | (0.01) | -0.59* | (0.01) |

Notes: *$p < .001$. β = standardized regression coefficient; *SE* = standard error.

# 7

# Gesamtdiskussion

# 7 Gesamtdiskussion

In dieser Arbeit wurde der Zusammenhang zwischen verschiedenen Komponenten der Testteilnahmemotivation und Testleistung analysiert. Ausgangspunkt bildet die Problematik, dass fehlende Motivation in Low-Stakes-Assessments eine valide Interpretation der Testergebnisse gefährden kann (Asseburg, 2011; Eklöf, 2007, 2008, 2010a, 2010b; Thelk, Sundre, Horst & Finney, 2009). Vorherige Forschung basierte meist auf der Erwartung-Wert-Theorie (Wigfield & Eccles, 2000), jedoch wurde bisher kein theoretisches Modell der Testteilnahmemotivation konstituiert, das alle drei Komponenten (Erwartung, Wert und Anstrengung) beinhaltet. Daher wurde im Theorieteil ein Erwartung-Wert-Anstrengung-Modell der Testteilnahmemotivation aufgestellt, das in drei Studien ausschnittsweise empirisch überprüft wurde. Die erste Studie konzentrierte sich hauptsächlich auf die Frage, ob die situationsspezifische Testteilnahmemotivation überhaupt einen Einfluss auf die gezeigte Testleistung aufweist, wenn für domänenspezifische motivationale Merkmale kontrolliert wird. Die zweite Studie untersuchte das komplexe Beziehungsgefüge von Erwartung, Wert, Anstrengung und Testleistung, wobei vor allem die Rolle von Anstrengungsbereitschaft als Mediatorvariable im Zentrum stand. Mit der dritten Studie wurde die Dynamik der motivationalen Prozesse im Verlauf einer Testsitzung in den Fokus der Betrachtungen gerückt.

Im Folgenden werden die wichtigsten Ergebnisse der drei Studien zusammengefasst (Abschnitt 7.1) sowie die Implikationen der Ergebnisse für das Erwartung-Wert-Anstrengung-Modell (Abschnitt 7.2) abgeleitet. Anschließend werden die Grenzen der vorliegenden Arbeit reflektiert (Abschnitt 7.3) und Konsequenzen für die Berichterstattung und die Assessment-Praxis (Abschnitt 7.4) geschlussfolgert. Abschließend wird ein Ausblick für zukünftige Forschungsthemen zur Testteilnahmemotivation skizziert (Abschnitt 7.5).

## 7.1 Zusammenfassung der Ergebnisse

### 7.1.1 Studie I

Um das komplexe Zusammenspiel der verschiedenen Testteilnahmekomponenten und Testleistung erforschen zu können, muss zunächst geklärt sein, ob die situationsspezifischen Motivationskomponenten überhaupt relevant für die Testleistung in Large-Scale-Assessments sind, wenn domänenspezifische motivationale Merkmale (z. B. Selbstkon-

zept) berücksichtigt werden (Fragestellung 1). Ebenfalls wurde der Zusammenhang zwischen Anstrengungsbereitschaft und verschiedenen Wertaspekten analysiert, da Anstrengungsbereitschaft als Ergebnis von Erwartung und Wert eine Sonderrolle in der Anwendung des Erwartung-Wert-Anstrengung-Modells für den Low-Stakes-Assessment-Kontext einnimmt, wie in Abschnitt 2.4.4 formuliert (Fragestellung 2). Mit besonderer Berücksichtigung des gegliederten Schulsystems in Deutschland, wurden beide Fragestellungen differenziert für Schülerinnen und Schüler an Gymnasien und an nichtgymnasialen Schularten betrachtet. Die Stichprobe bildeten Neuntklässlerinnen und Neuntklässler, die an der ersten PISA-Erhebung im Jahr 2000 teilnahmen und nach dem Mathematiktest auch Fragen zu ihrem Selbstkonzept in Mathematik als domänenspezifische Kompetenzüberzeugungen und Fragen zu ihrer situationsspezifischen Testteilnahmemotivation (Testattraktivität und -nützlichkeit, emotionale Befindlichkeit sowie Sorgen und Ablenkung vom Test) beantworteten.

Zusammenfassend ergaben die Analysen zur Beantwortung der ersten Fragestellung, dass nach Kontrolle der domänenspezifischen Kompetenzüberzeugungen die situationsspezifische Testteilnahmemotivation einen, wenn auch kleinen Zusammenhang mit der Testleistung aufwies. Aufgrund der widersprüchlichen Ergebnisse bisheriger Studien (Eklöf, 2007, 2008) wurden keine Annahmen formuliert. Insgesamt konnten 15 Prozent der Leistungsvarianz aufgeklärt werden, wobei circa die Hälfte auf die domänenspezifischen Kompetenzüberzeugungen, dem Selbstkonzept zurückging und die verbleibende Hälfte auf die verschiedenen Skalen der Testteilnahmemotivation. Teilnehmende mit einem höheren Selbstkonzept in Mathematik und einer höheren berichteten Anstrengungsbereitschaft schnitten besser ab als Teilnehmende mit einer niedrigerer Ausprägung auf diesen Variablen. Die Schülerinnen und Schüler, die sich während des Tests weniger Sorgen über ihre Fähigkeiten machten und sich weniger mit Test irrelevanten Dingen beschäftigten, zeigten eine höhere Testleistung als Teilnehmende, die mehr Sorgen und Ablenkung berichteten. Damit korrespondieren die Ergebnisse mit den Befunden von Eklöf (2008), aber widersprechen den Befunden von Eklöf (2007), die keinen signifikanten Zusammenhang zwischen Testteilnahmemotivation und Leistung nach Kontrolle domänenspezifischer Kompetenzüberzeugungen ergaben. Werden die Ergebnisse nach der Schulart differenziert betrachtet, konnten vor allem für die domänenspezifischen Kompetenzüberzeugungen Unterschiede festgestellt werden: Für Schülerinnen und Schüler an Gymnasien wies das Selbstkonzept in Mathematik einen stärkeren Zusammenhang mit der Mathematikleistung

auf als für Schülerinnen und Schüler an anderen Schularten. Ein weiteres Ergebnis ist, dass die Jugendlichen an nicht-gymnasialen Schularten ein besseres Testergebnis zeigten, je weniger sie während der Testbearbeitung abgelenkt waren.

Bezüglich des Zusammenhangs zwischen der berichteten Anstrengungsbereitschaft und den anderen situationsspezifischen motivationalen Aspekten ergaben die Analysen, dass Schülerinnen und Schüler, die den Test als attraktiv und nützlich wahrgenommen haben und diesen ablenkungsfrei und konzentriert bearbeiteten, eine hohe Anstrengungsbereitschaft berichteten. Damit wurden die Annahmen, dass vor allem der Interessenaspekt sowie die Kosten der Wertkomponente die Unterschiede in der investierten Anstrengungsbereitschaft erklären und nicht der Nützlichkeitsaspekt teilweise bestätigt. Der Interessensaspekt und die Kosten sowie der Nützlichkeitsaspekt der Wertkomponente tragen zur Erklärung der Anstrengungsbereitschaft bei. Ein Blick auf die schulartspezifischen Ergebnisse zeigte vor allem Unterschiede bezüglich der Rolle der Testattraktivität auf: Für Schülerinnen und Schüler an nicht-gymnasialen Schularten scheint ein attraktiver Test relevanter zu sein als für Gymnasiastinnen und Gymnasiasten. Ein Test, dessen Bearbeitung Freude bereitet, begünstigt für Lernende an nicht-gymnasialen Schularten die Bereitschaft, Anstrengung zu investieren.

### 7.1.2 Studie II

Die Ergebnisse der ersten Studie haben verdeutlicht, dass ein Zusammenhang zwischen Anstrengungsbereitschaft, Wert und Leistung in einem Low-Stakes-Test vorhanden ist. Auf Basis dieser Ergebnisse sowie aufgrund der Tatsache, dass die Erwartungskomponente in der bisherigen Testteilnahmemotivationsforschung vernachlässigt wurde, obwohl das Erwartung-Wert-Modell als theoretische Basis angewendet wurde, fokussierte die zweite Studie auf das komplexe Zusammenspiel zwischen Erwartung, Wert, Anstrengungsbereitschaft und Leistung. Ebenfalls wurde geprüft, ob sich die Beziehung zwischen den genannten Komponenten ändert, wenn Testteilnahmemotivation vor dem Leistungstest beziehungsweise nach dem Test erhoben wurde. Vor allem die Untersuchung der bisher vernachlässigten Zusammenhänge der Erwartungskomponente mit Anstrengungsbereitschaft und Leistung trägt zum theoretischen Erkenntnisgewinn bei. Datengrundlage hierfür bildete der IQB-Ländervergleich aus dem Jahr 2012, der unter anderem die Mathematikleistung von Neuntklässlerinnen und Neuntklässlern sowie deren Anstrengungsbereitschaft und Erfolgserwartungen erfasste. Überdies wurden die Wichtigkeit und das Interesse am Test sowie Misserfolgsbefürchtungen (Kosten) als Teil der Wertkomponente erhoben.

Die Ergebnisse dieser Studie bestätigen die in der Fragestellung (s. Abschnitt 3) aufgestellten Annahmen: Die Anstrengungsbereitschaft der Schülerinnen und Schüler konnte zu beiden Messzeitpunkten in hohem Maße durch den Wert erklärt werden, den sie dem Test beimaßen, aber auch zum Teil durch die Erfolgserwartung. Vor allem die wahrgenommene Wichtigkeit des Tests als Teil der Wertkomponente wies einen starken Zusammenhang mit der Anstrengungsbereitschaft auf. Teilnehmende wiederum, die eine höhere Anstrengungsbereitschaft sowie höhere Erfolgserwartungen berichteten, wiesen eine höhere Testleistung auf als Teilnehmende mit niedriger Ausprägung der Anstrengungsbereitschaft und Erfolgserwartung. Damit zeigte die zweite Studie insgesamt die Relevanz der wahrgenommen Erfolgserwartungen neben der Anstrengungsbereitschaft und Wertkomponente für die gezeigte Testleistung. Der Einfluss der wahrgenommenen Wichtigkeit des Tests auf die Testleistung wurde vollständig durch Anstrengungsbereitschaft vermittelt, womit auch die Bedeutsamkeit der Wertkomponente für eine hohe Testleistung aufgezeigt wurde. Interesse am Test und Erfolgserwartung zeigten nur sehr kleine indirekte Effekte auf Testleistung via Anstrengungsbereitschaft.

Die Jugendlichen wurden sowohl vor als auch nach dem Test zu ihrer Testteilnahmemotivation befragt. Die Vorhersage der Testleistung durch Erwartung, Wert und Anstrengung, die vor dem Test berichtet wurde, unterschied sich hinsichtlich der Vorhersage mit den drei Komponenten, die nach dem Test berichtet wurden, lediglich bezüglich der Vorhersagekraft der Erfolgserwartung: Die nach dem Test wahrgenommene Erfolgserwartung zeigte einen stärkeren Zusammenhang mit der Testleistung als die vor dem Test wahrgenommene Erfolgserwartung. Möglicherweise schätzen die Schülerinnen und Schüler ihre Erfolgserwartung nach dem Test realistischer ein, nachdem sie die Aufgaben bearbeitet haben, als vor dem Test, nachdem sie lediglich Beispielaufgaben gesehen haben.

### 7.1.3 Studie III

Sowohl im Grundmodell der klassischen Motivationspsychologie (Rheinberg, 2008) als auch im *demands-capacity model of test-taking effort* (Wise & Smith, 2011) wird auf eine globale Dynamik motivationaler Prozesse und auch speziell während einer Testsitzung hingewiesen. Daher fokussierte der letzte Artikel für diese Dissertation auf die Veränderung der Testteilnahmemotivation während der Bearbeitung eines zweistündigen Leistungstests sowie auf den Zusammenhang dieser Veränderung mit der gezeigten Testleistung. Dabei wurden die Erkenntnisse der ersten beiden Studien berücksichtigt und

es wurde das komplexe Beziehungsgefüge von Erwartung, Wert, Anstrengung und Testleistung modelliert. Dabei wurde das Selbstkonzept als Prädiktor der Erfolgserwartungen sowie folgende Kontrollvariablen berücksichtigt: domänenspezifische Kompetenzüberzeugungen, fachspezifische Leistung und sozio-ökonomischer Hintergrund der Schülerinnen und Schüler. Außerdem wurde aufgrund der Ergebnisse der zweiten Studie hier nur die wahrgenommene Wichtigkeit des Tests als Wertkomponente modelliert. Für die dritte Studie wurden erneut die Daten des IQB-Ländervergleichs verwendet, in dem die Testteilnahmemotivation nicht nur vor und nach dem Test, sondern auch nach der ersten Testhälfte abgefragt wurde.

Die Analysen zur Beantwortung der ersten Fragestellung, ob sich die Erwartungs- und Wertkomponente sowie die Anstrengungsbereitschaft während der Testsitzung verändern, ergaben, dass die Testteilnehmenden im Durchschnitt zunächst den Test selbstsicher mit hohen Erfolgserwartungen begannen, ihn als wichtig empfanden und bereit waren, Anstrengung zur Bearbeitung der Aufgaben zu investieren. Während der Testsitzung blieben die Schülerinnen und Schüler im Durchschnitt hinsichtlich ihrer Erfolgserwartung zuversichtlich, jedoch war eine moderate durchschnittliche Abnahme in der Wichtigkeit und in der Anstrengungsbereitschaft zu verzeichnen. Für Teilnehmende, die zu Beginn der Testsitzung eher bereit waren, sich bei den Aufgaben anzustrengen als der Durchschnitt, fiel die Anstrengungsbereitschaft im Verlaufe des Tests stärker ab als für Teilnehmende, die zu Beginn eine niedrigere Anstrengungsbereitschaft berichteten. Damit konnten zwei (stabiler Verlauf der Erfolgserwartungen und Abnahme in der Anstrengungsbereitschaft) der drei im Voraus formulierten Annahmen bestätigt werden (s. Abschnitt 3). Der postulierte stabile Verlauf der Wichtigkeit des Tests wurde durch die berichtete Abnahme in der Wichtigkeit nicht bestätigt.

In einem zweiten Schritt wurde untersucht, ob die Verläufe von Erwartung, Wert und Anstrengung zusammenhängen. Es konnte ein starker positiver Zusammenhang zwischen der Veränderung in der wahrgenommenen Wichtigkeit und der Veränderung in der Anstrengungsbereitschaft konstatiert werden. Außerdem zeigte, wenn auch schwächer als die eben beschriebenen Zusammenhänge, die Veränderung in der Erfolgswahrscheinlichkeit einen positiven Zusammenhang sowohl mit der Veränderung in der wahrgenommenen Wichtigkeit des Tests als auch mit der Veränderung in der Anstrengungsbereitschaft. Damit wurde die Annahme bestätigt, dass die Veränderungen in den drei Komponenten der

Testteilnahmemotivation zusammenhängen, was die Dynamik der motivationalen Prozesse unterstützt, die während der Bearbeitung eines Low-Stakes-Tests ablaufen.

Für die Vorhersage der gezeigten Testleistung als letzten Analyseschritt waren vor allem die anfängliche Anstrengungsbereitschaft und Erfolgserwartungen sowie die Veränderung in den Erfolgserwartungen von Bedeutung. Die anfängliche Wichtigkeit des Tests wirkte wie in Studie II indirekt via Anstrengungsbereitschaft vor dem Test auf die Testleistung. Obwohl die Erfolgserwartungen über alle Schülerinnen und Schüler hinweg einen stabilen Verlauf zeigten, war die interindividuelle Variabilität in den intraindividuellen Verläufen der situativen Erfolgserwartung während der Testsitzung hoch genug, dass trotz des durchschnittlichen stabilen Verlaufs ein Zusammenhang mit der Mathematikleistung nachweisbar war.

Neben der Beziehung zwischen der Veränderung in der Testteilnahmemotivation und der Testleistung wurden auf Anregung eines Gutachtens bezüglich Studie III ergänzende Berechnungen durchgeführt, die über das Manuskript hinausgehen. Da diese zum Verständnis der motivationalen Prozesse beitragen, sollen sie hier kurz berichtet werden. In einer Zusatzanalyse wurde der Zusammenhang zwischen der Veränderung in der Testteilnahmemotivation und der *Veränderung in der Testleistung* untersucht. Dafür wurde das Testergebnis der ersten Testhälfte als anfängliche Testleistung modelliert und die Veränderung in der Testleistung während der zweiten Testhälfte geschätzt. Diese Analyse ergab eine geringe Abnahme in der Testleistung während der zweiten Testhälfte, aber auch eine signifikante Variabilität in dieser Abnahme der Testleistung zwischen den Teilnehmenden. Die Untersuchung des Zusammenhangs zwischen der Veränderung in der Testteilnahmemotivation und der Veränderung in der Testleistung ergab, dass die geringe durchschnittliche Abnahme in der Testleistung während der zweiten Testhälfte weder mit der anfänglichen Erfolgserwartung, anfänglich wahrgenommenen Wichtigkeit des Tests oder anfänglich berichteten Anstrengungsbereitschaft zusammenhing noch mit der Veränderung in der Erfolgserwartung, der Veränderung in der wahrgenommenen Wichtigkeit des Tests oder der Veränderung in der Anstrengungsbereitschaft. Das heißt, die Veränderung in der Testleistung ließ sich nicht durch die anfängliche Testteilnahmemotivation erklären sowie auch nicht durch die Veränderung in der Testteilnahmemotivation während des Tests. Daher wurden die ursprünglich berichteten Analysen beibehalten und nicht modifiziert.

## 7.2 Implikationen für das Erwartung-Wert-Anstrengung-Modell

Die vorliegende Arbeit trägt zum akademischen Diskurs über die Testteilnahmemotivation bei, da sie verschiedene situationsspezifische motivationale Komponenten unter Anwendung der Erwartung-Wert-Theorie der Leistungsmotivation (Wigfield & Eccles, 2000) in zwei *large-scale low-stakes* Testsituationen untersuchte. Auf dieser Basis wurde ein an die Besonderheiten der Testteilnahmemotivation angepasstes Erwartung-Wert-Anstrengung-Modell aufgestellt und in drei Studien überprüft. Im Folgenden werden die Befunde differenziert diskutiert unter Betrachtung a) der Relation zwischen den motivationalen Komponenten des Erwartung-Wert-Anstrengung-Modells der Testteilnahmemotivation, nämlich Erwartung, Wert und Anstrengung und b) der Beziehungen zwischen diesen drei motivationalen Komponenten und Testleistung. Abschließend wird der Zusammenhang zwischen der Veränderung in den drei motivationalen Komponenten und Testleistung diskutiert.

### 7.2.1 Beziehung zwischen Erwartung, Wert und Anstrengungsbereitschaft

In Studie I wurde unter anderem Anstrengungsbereitschaft mit verschiedenen Aspekten der Wertkomponente in Beziehung gesetzt. Die Analysen ergaben, dass die Aspekte Interesse, Nützlichkeit und Kosten der Wertkomponente die investierte Anstrengungsbereitschaft vorhersagen. Dies stimmt mit den Ergebnissen von Cole et al. (2008) überein, die den Zusammenhang zwischen Interesse, Nützlichkeit, Wichtigkeit, Anstrengung und Testleistung auf Basis einer studentischen Stichprobe untersuchten. In ihrer Studie traten Testnützlichkeit und Testwichtigkeit als stärkste Prädiktoren von Anstrengungsbereitschaft hervor. Allerdings wurden in der ersten PISA-Studie, die die Datengrundlage für Studie I bildet, den Schülerinnen und Schülern keine Fragen zu ihren Erfolgserwartungen und ihrer Einschätzung der Wichtigkeit des Tests gestellt, so dass diese zwei Komponenten nicht untersucht werden konnten. Studie II berücksichtigte diese beiden Komponenten zusätzlich zum Interessenaspekt und Kostenaspekt der Wertkomponente und zur Anstrengungsbereitschaft, um die Beziehungen zwischen Erwartung, Wert und Anstrengung differenzierter untersuchen zu können.

Die Vorhersage der Anstrengungsbereitschaft durch Erwartung und Wert in Studie II zeigte, dass Anstrengungsbereitschaft vor allem durch die Wertkomponente vorhergesagt werden kann. Dabei stach die wahrgenommene Wichtigkeit als stärkster Prädiktor der Anstrengungsbereitschaft hervor. Dies korrespondiert mit den Ergebnissen bisheriger

Studien, die entweder hohe Korrelationen zwischen Wichtigkeit und Anstrengung feststellten (Abdelfattah, 2010; Barry & Finney, im Druck; Eklöf & Nyroos, 2013; Thelk et al., 2009) oder die unter den verschiedenen Aspekten der Wertkomponente die Wichtigkeit des Tests als stärksten Prädiktor von Anstrengung identifizierten (Cole et al., 2008; Knekta & Eklöf, im Druck). Knekta und Eklöf (im Druck) untersuchten die Beziehungen zwischen der berichteten Erfolgserwartung, Wichtigkeit, Interesse und Testangst von schwedischen Neuntklässlerinnen und Neuntklässlern. In ihrer Studie wurde außerdem ein starker Einfluss der Erfolgserwartungen auf die investierte Anstrengungsbereitschaft in einem Low-Stakes-Test aufgezeigt, den die Ergebnisse von Studie II in diesem Ausmaß nicht ergaben. Dies könnte zum einen an Unterschieden zwischen den untersuchten Fächern liegen (Mathematik in Studie II vs. die naturwissenschaftlichen Fächer bei Knekta & Eklöf). Möglicherweise sind die Zusammenhänge zwischen den Komponenten der Testteilnahmemotivation und Leistung vom getesteten Fach abhängig. Allerdings wurden bei Knekta und Eklöf (im Druck) die Berechnungen über alle drei naturwissenschaftlichen Fächer (Biologie, Chemie und Physik) hinweg vorgenommen und nicht pro Fach separiert, so dass Rückschlüsse auf einzelne Fächer nicht gezogen werden können. Zum Beispiel zeigten Cole et al. (2008) Unterschiede auf in den Fächern Englisch, Mathematik, Naturwissenschaften und Sozialwissenschaften bezüglich des Zusammenhangs verschiedener Skalen der Wertkomponente und Anstrengungsbereitschaft. Dabei variierte der Einfluss des Wichtigkeits- und Interessenaspekts der Wertkomponente auf die Anstrengungsbereitschaft zwischen diesen Fächern. In einer anderen schwedischen Studie fanden Eklöf und Nyroos (2013) hingegen nur geringfüge Unterschiede bezüglich der Korrelation von Wichtigkeit und Anstrengung beziehungsweise von Anstrengung und Testleistung zwischen den naturwissenschaftlichen Fächern Biologie, Chemie und Physik. Da aber in keiner der beiden letztgenannten Studien auch die Erwartungskomponente erhoben wurde, ist es denkbar, dass sich die Erfolgserwartung und deren Zusammenhänge mit Anstrengungsbereitschaft und Leistung zwischen den Fächern unterscheiden. Beispielsweise zeigte die Studie von Haag und Götz (2012), dass Schülerinnen und Schüler der achten und elften Klasse das Fach Mathematik als schwieriger und anstrengender einschätzen als die Fächer Physik und Biologie, aber eine gute Mathematiknote als wichtiger bewerteten als eine gute Physik- oder Biologienote. Möglicherweise spiegelt sich dieser Effekt auch in den Unterschieden in der Testteilnahmemotivation zwischen den Fächern wider.

Zum anderen könnten die divergierenden Ergebnisse auch kulturellen Unterschieden bezüglich der Erfolgserwartung in Low-Stakes-Tests geschuldet sein. Bisherige Studien fanden zwar nur geringe Unterschiede in der berichteten Anstrengung zwischen verschiedenen Ländern, wie beispielsweise die an PISA 2003 teilnehmenden Länder (Butler & Adams, 2007) oder zwischen den Ländern Norwegen, Schweden und Slowenien (Eklöf et al., 2014). Überdies war die Beziehung zwischen Anstrengungsbereitschaft und Testleistung über die Länder hinweg ähnlich ausgeprägt (Eklöf et al., 2014). Dennoch wurde hier erneut nicht die Erwartungs- oder die Wertkomponente in die Untersuchung mit einbezogen, so dass die These kultureller Unterschiede der Testteilnahmemotivation nicht gänzlich entkräftet werden kann. Selbst wenn in der westlichen Welt nicht allzu große kulturelle Unterschiede vermutet werden, so sind dennoch nationale Unterschiede in der Organisation und Gestaltung des Unterrichts denkbar, wie zum Beispiel die Dauer der Schulstunden oder der Leistungstests oder die zeitliche Nähe zu Leistungsüberprüfungen, die sich auf die Testteilnahmemotivation auswirken können.

Zusammenfassend scheint es, dass die Wertkomponente entscheidend für die Vorhersage der Anstrengungsbereitschaft ist, wie aufgrund der Theorie und Empirie zu erwarten war. Die Erfolgserwartungen sind ebenfalls von Bedeutung, jedoch ist der gefundene Zusammenhang mit Anstrengungsbereitschaft kleiner als die Beziehung zwischen der Wertkomponente und Anstrengungsbereitschaft. Die theoretische Konzeptualisierung der Anstrengungsbereitschaft als Resultat von Erwartung und Wert scheint sich auch in der Empirie zu bestätigen. Innerhalb der Wertkomponente zeigten die hier durchgeführten Studien, dass alle vier Wertaspekte, also Wichtigkeit, Interesse, Nützlichkeit und Kosten, Prädiktoren der Anstrengungsbereitschaft sind. In Studie I war vor allem der Interessenaspekt ein starker Prädiktor für Anstrengung; in Studie II war es vor allem der Wichtigkeitsaspekt der Wertkomponente. Die Annahme, dass der Nützlichkeitsaspekt in Low-Stakes-Assessments per se niedrig ist und keinen Zusammenhang mit Anstrengung aufweist, wurde nicht bestätigt, so dass die Berücksichtigung aller vier Wertaspekte auch in Low-Stakes-Testsituationen von Bedeutung ist.

### 7.2.2 Beziehung zwischen Erwartung, Wert, Anstrengungsbereitschaft und Leistung

Bevor im folgenden Abschnitt die drei Komponenten der Testteilnahmemotivation (Erwartung, Wert und Anstrengung) mit der gezeigten Testleistung in Verbindung

gebracht werden, wird diskutiert, ob Testteilnahmemotivation überhaupt einen Zusammenhang mit Testleistung aufweist, wenn domänenspezifische Kompetenzüberzeugungen berücksichtigt werden. Die Ergebnisse der Studie I auf Basis der Daten der ersten PISA-Erhebung ergaben, dass auch unter Kontrolle des domänenspezifischen Selbstkonzepts, die Testteilnahmemotivation einen, wenn auch schwachen, Zusammenhang mit Testleistung aufwies. Damit widersprechen diese Befunde einerseits denen von Eklöf (2007), die keinen signifikanten Zusammenhang von Testteilnahmemotivation und Leistung nach Kontrolle domänenspezifischer Kompetenzüberzeugungen fand. Anderseits unterstützen sie die Ergebnisse von Eklöf (2008), die einen signifikanten, wenn auch schwachen, Zusammenhang zwischen Testteilnahmemotivation und Testleistung in der schwedischen TIMS-Studie ergaben, nachdem für domänenspezifische Kompetenzüberzeugungen kontrolliert wurde. Möglicherweise ist dies auf die Verwendung unterschiedlicher Personenschätzer zurückzuführen. Sowohl in Eklöfs Studie aus dem 2008 als auch in Studie I wurden als Personenschätzer fünf *Plausible Values* verwendet, die aus der Skalierung des Leistungstests unter Verwendung der probabilistischen Testtheorie gewonnen wurden (Yen & Fitzpatrick, 2006). In Eklöfs Studie aus dem Jahr 2007 wurde berichtet, dass ein „national mathematics Rasch score“ (S. 317) als Personenschätzer eingesetzt wurde. Ob dieser Schätzer nun einen einzelnen *Plausible Value* darstellte oder ein anderer Personenschätzer der Berechnung zugrunde lag, kann leider nicht geklärt werden.

Die zweite Studie auf Grundlage neuer Daten des IQB-Ländervergleichs ergab, dass vor allem die Erfolgserwartung und die Anstrengungsbereitschaft, die nach dem Test berichtet wurden, starke Prädiktoren der Testleistung darstellen. Der Zusammenhang zwischen der Erfolgserwartung und Testleistung war dabei stärker als der Zusammenhang zwischen der Anstrengungsbereitschaft und Testleistung. Damit bestätigen die Analysen die Ergebnisse der Studie von Asseburg (2011) und Freund und Holling (2011), die die Erfolgserwartung als starken Prädiktor von Testleistung ergaben. Die letztgenannte Studie führte ein bis zwei Wochen nach der ersten Erhebung einen Retest ihres Intelligenztests durch. Die Vorhersage der Retest-Leistung zeigte einen höheren Zusammenhang zwischen der Erfolgserwartung und der Testleistung, aber auch zwischen den Aspekten der Wertkomponente und der Testleistung (Freund & Holling, 2011). Dies legt nahe, dass durch die steigende Erfahrung, die die Teilnehmenden während des Tests sammeln, sich auch die Vorhersagekraft der Erfolgserwartung für die Testleistung erhöht. Damit wird erneut deutlich, dass die in Abschnitt 2.4.3 benannten Studien, die die Erwartungskomponente

vernachlässigen, eine bedeutende Komponente des Erwartung-Wert-Anstrengung-Modells auslassen. Die Argumentation von Cole et al. (2008), dass diese Komponente nicht erhoben werden muss, weil die Teilnehmenden ihren Erfolg nicht einschätzen können, kann nicht bestätigt werden. Obwohl die Testeilnehmenden keine direkte Rückmeldung zu ihrem Testergebnis erhalten, scheint die Erfolgserwartung der Schülerinnen und Schüler und damit ihre innere Zuversicht ein wichtiger Motor für die Testteilnahmemotivation zu sein, der sich wiederum auf die Testleistung auswirkt.

Der durch eine Vielzahl internationaler Studien (Eklöf & Nyroos, 2013; Knekta & Eklöf, im Druck; Thelk et al., 2009; Wise & DeMars, 2005) bestätigte Zusammenhang zwischen Anstrengungsbereitschaft und Testleistung in Low-Stakes-Assessments konnte in Studie I und II repliziert werden. Darüber hinaus korrespondieren die Ergebnisse von Studie II mit den Befunden von Cole et al. (2008), dass die Anstrengungsbereitschaft den Zusammenhang zwischen den verschiedenen Aspekten der Wertkomponente und Leistung mediiert. Generell konnte kein direkter Zusammenhang zwischen der Testleistung und der Wertkomponente in Studie II festgestellt werden. Allerdings wurde der Einfluss der verschiedenen Wertaspekte auf die Testleistung mindestens partiell, für die Wichtigkeit sogar vollständig über die Anstrengungsbereitschaft vermittelt. Dies könnte eine Erklärung für die Ergebnisse der Studie von Asseburg (2011) sein. Ihre Untersuchung zum Einfluss der Erwartungs- und Wertkomponente auf Testleistung mit Neuntklässlerinnen und Neuntklässlern (s. Abschnitt 2.4.3) ergab, dass lediglich die Erfolgserwartungen die Testleistung vorhersagten und nicht die Wertkomponente. Dies ist möglicherweise darauf zurückzuführen, dass die investierte Anstrengung der Testteilnehmenden nicht erhoben wurde. Ohne die Erfassung der Anstrengungsbereitschaft konnten keine indirekten Effekte der Wertkomponente auf die Testleistung modelliert werden. Andere Studien hingegen, die allerdings auch keine Anstrengungsbereitschaft erhoben, fanden einen direkten Zusammenhang zwischen dem Interessenaspekts der Wertkomponente und der Testleistung (Freund & Holling, 2011; Freund et al., 2011). Wie im Erwartung-Wert-Anstrengung-Modell postuliert, sollte auf Basis der Ergebnisse von Studie II die Anstrengungsbereitschaft der Testteilnehmenden in Untersuchungen im Kontext von Testteilnahmemotivation berücksichtigt werden, da diese sowohl einen starken Prädiktor von Testleistung als auch eine bedeutende Mediatorvariable für den Effekt der Wertkomponente auf Testleistung darstellt.

Insgesamt konnte in der vorliegenden Arbeit das theoretisch postulierte Erwartung-Wert-Anstrengung-Modell fast vollständig bestätigt werden: Sowohl die Erwartungskomponente als auch die Wertkomponente sagten die Anstrengungsbereitschaft der Schülerinnen und Schüler vorher; die Anstrengungsbereitschaft und die Erwartungskomponente sind wiederum prädiktiv für die Testleistung. Bezüglich des Zusammenhangs zwischen den verschiedenen Aspekten der Wertkomponente und Testleistung divergieren die Ergebnisse leicht: In Studie I wurde ein direkter Effekt der Kostenaspekt auf die Testleistung gefunden; in Studie II wurde der Effekt für den Wichtigkeitsaspekt vollständig und für den Interessen- und Kostenaspekt auf Testleistung partiell über die Anstrengungsbereitschaft mediiert. Ob die Beziehung zwischen Testnützlichkeit und Testleistung ebenfalls über die Anstrengungsbereitschaft vermittelt ist, konnte in Studie II nicht untersucht werden aufgrund fehlender Fragen zur Nützlichkeit des Tests. Insgesamt kann geschlussfolgert werden, dass alle drei Aspekte der Testteilnahmemotivation, also die Erfolgserwartungen, der wahrgenommene Wert des Tests und die Anstrengungsbereitschaft erhoben werden sollten, um das komplexe Zusammenspiel der Konstrukte untereinander und deren Zusammenhang mit Testleistung modellieren zu können.

*Beziehung zwischen dem Kostenaspekt der Wertkomponente und Testleistung*

Dem Kostenaspekt in der Erwartung-Wert-Theorie (Wigfield & Eccles, 2000) wurde in der bisherigen Forschung wenig Beachtung geschenkt. Im Kontext der Testteilnahmemotivation wird die Testangst den Kosten zugeordnet. Testangst wird als situationsspezifische Disposition definiert, in Situationen mit erhöhter Angst zu reagieren, die mit Tests und Leistung verbunden sind. Dabei werden meist die kognitive Komponente (Sorgen) und die affektive Komponente (Emotionalität) voneinander unterschieden (Hodapp, Glanzmann & Laux, 1995), die jedoch aufgrund des Zusammenhangs zwischen den beiden Komponenten meist kombiniert erhoben werden (Nie, Lau & Liau, 2011). Bei der Teilnahme an Low-Stakes-Tests sollten theoretisch die Schülerinnen und Schüler im Durchschnitt eine geringere Testangst im Vergleich zu High-Stakes-Tests berichten, da auch ein komplettes Versagen bei der Bearbeitung der Aufgaben keine negativen Folgen mit sich bringt. Daher wird in diesem Abschnitt der Kostenaspekt explizit diskutiert.

In Studie I zeigten die Sorgen, die sich auf die kognitive Komponente der Testangst beziehen, einen relativ starken Zusammenhang mit der gezeigten Testleistung. Dieses Ergebnis korrespondiert mit den Ergebnissen von Baumert und Demmrich (2001), die in der Pilotierungsstudie zur ersten PISA-Erhebung ebenfalls Anstrengungsbereitschaft und

Sorgen bezüglich der eigenen Fähigkeiten als stärkste Prädiktoren der Testleistung fanden. Allerdings klärten diese beiden Variablen in der PISA-Pilotierungsstudie fast doppelt so viel Varianz in der Testleistung auf, als die Analysen in Studie I mit allen domänen- und situationsspezifischen Prädiktoren erklären konnten. Ob dieser extreme Unterschied an dem kürzeren Leistungstest in der Pilotierungsstudie lag, kann hier nur vermutet werden, da keine Dokumentation der Analyse vorliegt und lediglich die aufgeklärte Varianz berichtet wird (Baumert & Demmrich, 2001, S. 457). Studie II konnte diesen Einfluss der Misserfolgsbefürchtungen, die eher die affektive Komponente der Testangst erfassen, auf die Testleistung allerdings nicht bestätigen. Auch Eklöf und Nyroos (2013) fanden lediglich einen schwachen Zusammenhang zwischen Testangst und Leistung sowie gar keinen zwischen Testangst und Anstrengungsbereitschaft. Die unterschiedlichen Befunde hinsichtlich des Zusammenhangs zwischen dem Kostenaspekt und der Testleistung in Studie I und II können zum einen der unterschiedlichen Operationalisierung geschuldet sein (Sorgen in Studie I vs. Misserfolgsbefürchtungen in Studie II). Zum anderen fehlt in Studie I das Konstrukt der Erfolgserwartungen. Detaillierte Analysen in Studie II ergaben, dass, wenn nur die Aspekte der Wertkomponenten und Anstrengungsbereitschaft als Prädiktoren der Testleistung verwendet wurden, die Misserfolgsbefürchtungen einen schwachen, aber signifikanten Einfluss auf Testleistung aufwiesen. Dieser Effekt lag in derselben Größenordnung wie der Einfluss der Beschäftigung mit Test irrelevanten Dingen auf die Testleistung in Studie I. Werden die Erfolgserwartungen als zusätzlicher Prädiktor mit in die Analyse einbezogen, verschwindet der Zusammenhang zwischen Misserfolgsbefürchtungen und Leistung in Studie II. Aufgrund der unterschiedlichen Konstrukte, die den beiden Studien zugrunde lagen, und den divergieren Ergebnissen, kann keine Schlussfolgerung bezüglich des Zusammenhangs zwischen dem Kostenaspekt und Testleistung gezogen werden.

Andere Studien, die auf die Erklärung der Unterschiede in der Testangst fokussierten anstatt auf die Erklärung von Unterschieden in der Leistung, zeigten ebenfalls komplexe Beziehungsgefüge zwischen Selbstwirksamkeit, Aufgabenwichtigkeit und Testangst (Nie et al., 2011) sowie zwischen Selbstkonzept, Zielorientierungen und Testangst (Putwain & Daniels, 2010). Damit wird das Potential andere Motivationstheorien zu domänenspezifischen Erwartungen (z. B. Selbstkonzept, s. Abschnitt 2.2.1) und Werten (z. B. Zielorientierungen, s. Abschnitt 2.2.2) deutlich, so dass auch hier weiterführende Analysen zur Erklärung der Testangst in Low-Stakes-Assessments vielversprechend sind.

### 7.2.3 Veränderung der Testteilnahmemotivation und die Beziehung mit Testleistung

Die dritte Studie dieser Arbeit konzentrierte sich auf die Veränderung in der Testteilnahmemotivation und den Zusammenhang der motivationalen Veränderung mit der tatsächlich gezeigten Testleistung. Die empirischen Beziehungen des getesteten Modells sind in Abbildung 7.1 dargestellt. Dabei sind in der Abbildung die Veränderungen in Erwartung, Wert und Anstrengung erneut mit einem grauen Kästchen gekennzeichnet, das eine Uhr enthält. Die vordere Ebene von Erwartung, Wert und Anstrengungsbereitschaft stellt die anfängliche Ausprägung der drei Konstrukte vor der Bearbeitung des Leistungstests dar.

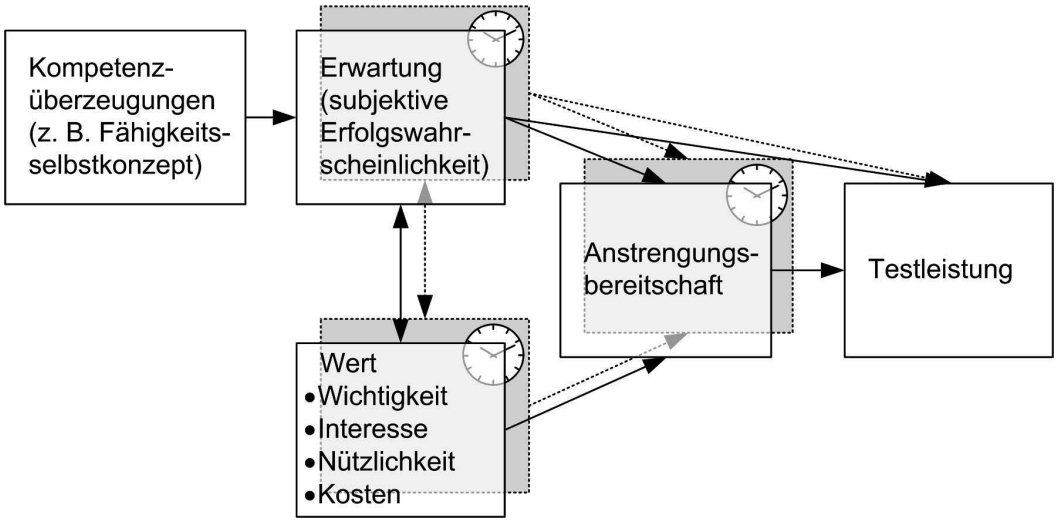



*Abbildung 7.1*. Empirische Zusammenhänge zwischen den Kompetenzüberzeugungen, der anfänglichen Ausprägung von Erwartung, Wert und Anstrengungsbereitschaft sowie deren Veränderungen und Testleistung (der Zusammenhang zwischen der Veränderung des Wertes und der Veränderung Anstrengungsbereitschaft wurde nur für den Wichtigkeitsaspekt untersucht).

Die Untersuchung der Veränderung in den Erfolgserwartungen, der Wichtigkeit des Tests und der Anstrengungsbereitschaft ergab im Durchschnitt einen relativ stabilen Verlauf der Erfolgserwartungen während der zweistündigen Testsitzung. Allerdings konnte auch Variabilität in dieser Veränderung der Erfolgserwartungen aufgezeigt werden, das heißt, es ergaben sich auch individuelle ansteigende sowie sinkende Verläufe. Da bis zu diesen Zeitpunkt keine weitere Studie die Veränderungen in den Erfolgserwartungen erhoben hat, kann dieses Ergebnis momentan nicht in die bisherige Literatur eingeordnet werden. Zumindest kann vermutet werden, dass die Schülerinnen und Schüler im Durch-

schnitt ihre Erfolgswahrscheinlichkeit für den Mathematiktest schon vor dem Test realistisch einschätzten und diese nicht korrigierten. Im Gegensatz dazu ergaben sich sinkende Verläufe für die Anstrengungsbereitschaft sowie für die Wichtigkeit des Tests. Die Abnahme in der Anstrengungsbereitschaft bestätigt bisherige Studien (Cao & Stokes, 2008; Horst, 2010), die eine abnehmende Anstrengungsbereitschaft innerhalb eines Leistungstests beobachten konnten. Allerdings stimmen sie nicht mit dem Ergebnis von Barry und Finney (im Druck) überein, die einen leichten Anstieg in der berichteten Anstrengung feststellten. Dieser Anstieg lässt sich wahrscheinlich darauf zurückführen, dass die letztgenannte Studie die Anstrengung über mehrere verschiedene Tests (ein Leistungstests und vier Einstellungstests) abfragte, während der Ländervergleich auf die Veränderung in der Testteilnahmemotivation innerhalb *eines* Leistungstests fokussierte. Für die wahrgenommene Wichtigkeit wurde basierend auf den Ergebnissen von Horst (2010) ein stabiler Verlauf angenommen, der sich aufgrund des abnehmenden Verlaufs in den Daten empirisch nicht bestätigte. Möglicherweise ist dies der unterschiedlichen Operationalisierung geschuldet: Im Gegensatz zu Horst (2010), die die Wichtigkeit des 75-minütigen Tests direkt erfragte, wurde im Ländervergleich die Wichtigkeit des zweistündigen Tests indirekt mithilfe des Konstrukts der Herausforderung erhoben (s. Abschnitt 7.3). Es ist allerdings ebenso denkbar, dass aufgrund der längeren Bearbeitungszeit im Ländervergleich der leistungsthematische Anreiz des Tests abgenommen hat.

Die Veränderungen in den Konstrukten der Testteilnahmemotivation wiesen alle positive Zusammenhänge miteinander auf: Die Veränderung in der Wichtigkeit des Tests ist mit der Veränderung in der Anstrengungsbereitschaft verbunden; Veränderung in den Erfolgserwartungen hängen, wenn auch nicht ganz so stark wie die Wichtigkeit, mit der Veränderung in der Anstrengungsbereitschaft sowie mit der Veränderung in der Wichtigkeit des Tests zusammen. Zum aktuellen Zeitpunkt ist bisher wenig Forschung zur Veränderung in der Testteilnahmemotivation vorhanden. Barry und Finney (im Druck) fanden keinen Zusammenhang zwischen der Veränderung in der Anstrengung und Veränderung in der Wichtigkeit des Tests. Dies könnte darin begründet sein, dass die zuletzt erwähnte Studie den Verlauf während einer Testbatterie mit einem Leistungstest und vier Einstellungstests erforschte und im Ländervergleich der Verlauf der Testteilnahmemotivation innerhalb eines Leistungstests von Interesse war. Die Befunde zeigen die Dynamik und das komplexe Zusammenspiel der einzelnen Komponenten der Testteilnahmemotivation, die innerhalb einer Testperson während einer Testsitzung ablaufen. Als

bedeutsam für die Vorhersage der Testleistung stellte sich die Veränderung in den Erfolgserwartungen heraus, was erneut untermauert, wie wichtig die Erhebung aller Komponenten der Testteilnahmemotivation in einem Low-Stakes-Test ist. Die Veränderung in der Wichtigkeit des Tests sowie in der Anstrengungsbereitschaft zeigte keine Assoziation mit Leistung, so dass das theoretisch aufgestellte Erwartung-Wert-Anstrengung-Modell nicht für den Zusammenhang zwischen Testleistung und der Veränderung in der Testteilnahmemotivation zu gelten scheint. Eine Kritik könnte an dieser Stelle sein, dass die Testleistung lediglich „statisch“ erfasst wurde und nicht die Veränderung in der Testleistung betrachtet wurde. Theoretisch scheint es plausibel, dass die Veränderung in der Testteilnahmemotivation mit einer Veränderung in der Testleistung zusammenhängt. Jedoch zeigten auch die zusätzlichen Analysen keine Zusammenhänge, die die *Veränderung* in der Testleistung mit der Veränderung in den Konstrukten der Testteilnahmemotivation in Relation setzten. Möglicherweise liegt es an der untersuchten Zeitspanne von einer Zeitstunde für die Veränderung in der Testleistung, die zu kurz sein könnte, um Zusammenhänge zwischen der Veränderung in der Testleistung und der Veränderung in der Testteilnahmemotivation aufdecken zu können. Weitere Studien, die auf die Dynamik der verschiedenen Komponenten der Testteilnahmemotivation während einer Testsitzung fokussieren, sollten Fragen zur Erwartung und Anstrengungsbereitschaft sowie zum Wert mehrfach stellen. Soll allerdings die gezeigte Testleistung mit den motivationalen Aspekten in Relation gesetzt werden, scheint eine einmalige Erhebung der Testteilnahmemotivation ausreichend.

Der aufgezeigte Zusammenhang zwischen der Veränderung in der Erfolgserwartung und der Gesamtleistung hingegen korrespondiert mit dem in Abschnitt 2.4.2 vorgestellten *demands-capacity model of test-taking effort* (Wise & Smith, 2011). Dieses Modell erklärt theoretisch neben der anfänglichen Anstrengung auch die Anstrengung im Verlauf eines Tests. Dabei beeinflusst unter anderem die Bearbeitung vorangehender Aufgaben in einem Testheft die Zuversicht, die restlichen Aufgaben erfolgreich zu lösen, und damit die aufgewendete Anstrengungsbereitschaft. Dies zeigte sich in Studie III auch empirisch. Anscheinend wirkt die Veränderung der wahrgenommenen Erfolgserwartungen auf die Veränderung in der Anstrengungsbereitschaft und auch direkt auf die gezeigte Leistung. Allerdings wurden keine theoretischen Beziehungen zwischen der Veränderung in der Testteilnahmemotivation und Testleistung im Modell von Wise und Smith (2011) aufgestellt. Für die Forschung wird demnach erneut die Relevanz der Erhebung der

Erfolgserwartungen für die Testleistung verdeutlicht. Bei Vernachlässigung der Erwartungskomponente wird Gefahr gelaufen, einen wichtigen Bestandteil der Testteilnahmemotivation nicht zu beachten. Die praktische Bedeutung der Erhebung der Erfolgserwartungen wird im in Abschnitt 7.4 besprochen.

## 7.3 Grenzen und Empfehlungen für weitere Forschung

Bevor neben den theoretischen Implikationen auch praktische Konsequenzen abgeleitet werden, sollten einige Einschränkungen der hier durchgeführten Studien beziehungsweise Empfehlungen für nachfolgende Studien genannt werden. Grundlegend ist als Einschränkung zu nennen, dass die vorliegende Arbeit die Testteilnahmemotivation mit Fragebögen erfasste, die auf Selbsteinschätzungen der Schülerinnen und Schüler basierten. Damit wird vorausgesetzt, dass die Schülerinnen und Schüler die Skala, auf der sie ihre Testteilnahmemotivation positionieren sollten, richtig interpretieren konnten und ihre Motivation ehrlich berichteten. Diese Voraussetzungen lassen sich jedoch nicht empirisch überprüfen, was allerdings für alle Studien gilt, die Konstrukte auf Basis von Selbsteinschätzungen der Testpersonen erfassen.

Die Erfassung der einzelnen Wertaspekte divergierte zwischen den Studien I und II, was möglicherweise zu den unterschiedlichen Ergebnissen bezüglich der Zusammenhänge zwischen einzelnen Wertaspekten und Testleistung führte. So wurde in Studie I der Interessenaspekt der Wertkomponente (d. h. die wahrgenommene Freude während der Testbearbeitung, s. Abschnitt 2.4.4) über die Testattraktivität operationalisiert, in Studie II dagegen durch das Interesse am Test. Dabei zielten die Fragen der Testattraktivitätsskala mehr auf die Begeisterung während der Testbearbeitung ab („Wie viel Spaß hat dir der Test gemacht?“), während die Interessenskala eher das allgemeine Interesse an Low-Stakes-Tests erfasste („Bei Tests wie diesem brauche ich keine Belohnung; sie machen mir auch so viel Spaß“). Beim Vergleich der Vorhersagekraft des unterschiedlich operationalisierten Interessenaspektes der Wertkomponente zeigte sich, dass Testattraktivität ein stärkerer Prädiktor für Anstrengungsbereitschaft war als das Interesse am Test. Der stärkere Zusammenhang zwischen Testattraktivität und Anstrengungsbereitschaft reflektiert auch die größere inhaltliche Nähe der Testattraktivität mit der Definition des Interessenaspekts der Testteilnahmemotivation, wie in Abschnitt 2.4.4 erläutert wurde. Wird eine Verbindung vom Interessenaspekt der Wertkomponente im Kontext der Testteilnahmemotivation mit dem Konstrukt des Interesses (Hidi & Harackiewicz, 2000; Schiefele, 1999) geschlagen, dann scheint der Interessensaspekt auch theoretisch näher am situativen

Interessensbegriff zu sein als am individuellen Interessensbegriff (s. Abschnitt 2.2.2). Individuelles Interesse ist eine relativ stabile motivationale Orientierung (Hidi & Harackiewicz, 2000), wie zum Beispiel das generelle Interesse an Low-Stakes-Tests (Studie II); situatives Interesse ist als ein emotionaler Zustand definiert, der durch die Tätigkeit hervorgerufen wird und von Begeisterung begleitet werden kann (Hidi & Harackiewicz, 2000; Schiefele, 1999), wie die hervorgerufene Freude bei der Bearbeitung eines Low-Stakes-Tests (Studie I). Aus der Perspektive der Erwartung-Wert-Theorie scheint daher die Testattraktivität näher am theoretischen und eher situativen Interessenaspekt des postulierten Erwartung-Wert-Anstrengung-Modells zu sein als das eher generelle Interesse an Low-Stakes-Tests. Weitere Forschung in diesem Bereich wird daher an dieser Stelle die Erfassung der Testattraktivität beziehungsweise der Freude empfohlen, die die Teilnehmenden durch die Bearbeitung des Tests erfahren.

Eine Limitation der zweiten und dritten Studie betrifft die Erfassung des Wichtigkeitsaspekts der Wertkomponente. Zum einen wurde die wahrgenommene Wichtigkeit des Tests nur indirekt mit der Herausforderungsforderungsskala des Fragebogens zur Erfassung der aktuellen Motivation erfasst (Freund et al., 2011; Rheinberg et al., 2001). Die Fragen zur Herausforderung ermitteln, ob die Schülerinnen und Schüler die Testsituation als Leistungssituation wahrgenommen haben. Wird die Testsituation als Leistungssituation anerkannt, dann sollen die Schülerinnen und Schüler diese auch erfolgreich meistern wollen. Dadurch erlangt der Test an Wichtigkeit für die Teilnehmenden (Vollmeyer & Rheinberg, 2006). In den meisten Studien wird die wahrgenommene Wichtigkeit des Tests direkt erhoben, weshalb für die internationale Anschlussfähigkeit die direkte Erfassung des Wichtigkeitsaspekts empfohlen wird.

Überdies muss eingeräumt werden, dass in keiner der drei Studien alle vier Aspekte der Wertkomponente zusammen untersucht wurden. In Studie I wurde die Wichtigkeit des Tests nicht erhoben und in Studie II und III fehlte die Nützlichkeit des Tests. Der wahrgenommene Wert einer Aufgabe beziehungsweise eines Tests wird aus Sicht der Erwartung-Wert-Theorie durch alle vier Wertaspekte gleichzeitig konstituiert. Ein und dieselbe Aufgabe beziehungsweise ein Test kann dabei mehrere Aspekte des Wertes besitzen und je mehr Aspekte aktiviert sind, desto höher sollte der Wert ausfallen (Eccles, 2005). Demnach könnte es durch die Nichtbetrachtung eines Wertaspektes möglich sein, dass der wahrgenommene Wert insgesamt unterschätzt wird. Allerdings ist es ebenfalls vorstellbar, dass in Low-Stakes-Assessments nicht alle Wertaspekte aktiviert sein müssen, damit die

Schülerinnen und Schüler motiviert sind. Die individuelle Nützlichkeit der Bearbeitung eines Tests ohne Konsequenzen, das heißt die Passung einer Testbearbeitung mit zukünftigen Zielen, ist möglicherweise per se sehr gering und daher im Kontext der Testteilnahmemotivation weniger relevant als die anderen Wertaspekte. Allerdings zeigte Studie I einen, wenn auch geringen, Zusammenhang zwischen Testnützlichkeit und Leistung sowie Anstrengungsbereitschaft. Jedoch wurde der Nützlichkeitsaspekt nur mit einer Frage erhoben. Daher sollte zukünftige Forschung unter Einsatz mehrerer Fragen zur wahrgenommenen Testnützlichkeit diese Zusammenhänge genauer beleuchten.

## 7.4 Konsequenzen für das Bildungsmonitoring und die Assessment-Praxis

Wie bereits in Abschnitt 7.2 erwähnt, wurde in dieser Arbeit zum ersten Mal Testteilnahmemotivation in einer realen Testsituation untersucht, in der Schülerinnen und Schüler der neunten Klasse einen zweistündigen *Paper-and-Pencil-Test* bearbeiteten und in der alle für Testteilnahmemotivation relevanten Komponenten der Erwartung-Wert-Theorie, das heißt Erwartung, Wert und Anstrengungsbereitschaft mit der tatsächlich gezeigten Testleistung in Verbindung gebracht wurden. Die untersuchte Testsituation kann als prototypisch für eine Vielzahl von großangelegten Schulleistungsstudien ohne Konsequenzen angesehen werden, die seit der Jahrtausendwende vermehrt in den Schulen in Deutschland Einzug gehalten haben. Bisherige nationale Studien, die sich mit der berichteten Motivation von Schülerinnen und Schülern der Sekundarstufe I in Low-Stakes-Tests beschäftigten, verwendeten eine (teilweise viel) kleinere Anzahl von Testaufgaben, als es in Schulleistungsstudien üblich ist (Asseburg, 2011; Baumert & Demmrich, 2001; Freund et al., 2011), und setzten zum Teil nur geschlossene Antwortformate ein (Asseburg, 2011; Freund et al., 2011). Eine steigende Testlänge geht mit erhöhter Müdigkeit einher, was wiederum mit der investierten Anstrengungsbereitschaft verbunden ist (Ackerman & Kanfer, 2009). Ebenfalls erweisen sich offene Antwortformate in der Praxis im Mittel als schwieriger als geschlossene Formate (Klieme, Baumert, Köller & Bos, 2000; Leucht, Harsch, Pant & Köller, 2012), was auch die Bereitschaft beeinflussen kann, Anstrengung zu investieren (Wise & Smith, 2011). Um die Testteilnahmemotivation in Large-Scale-Assessments abbilden zu können, sollte diese unter realen Testbedingungen bezüglich der Testlänge und eingesetzten Aufgabenformaten untersucht werden, was in der vorliegenden Arbeit zum einen in der ersten PISA-Erhebung als auch im IQB-Ländervergleich 2012 umgesetzt wurde.

Auch wenn die durchgeführten Studien dieser Arbeit vor allem zu dem wissenschaftlichen Diskurs zur Anpassung des Erwartung-Wert-Modells an die Spezifika der Testteilnahmemotivation einen Beitrag leisten, sind neben den besprochenen theoretischen Implikationen aus den Ergebnissen einige praktische Bedeutsamkeiten ableitbar. Die ersten beiden Implikationen betreffen die Fragen, ob Testteilnahmemotivation eine Einschränkung valider Interpretationen der Testergebnisse in Low-Stakes-Tests darstellt und zu welchem Zeitpunkt während der Testsitzung die Testteilnehmenden zu ihrer Testteilnahmemotivation befragt werden sollten. Anschließend folgt ein Abschnitt, der sich mit praktischen Implikationen der Ergebnisse zur Erwartungskomponente auseinandersetzt, in dem Erfolgserwartungen mit Aufgabenschwierigkeiten in Beziehung gesetzt werden. Danach werden Implikationen der Ergebnisse zur Wertkomponente benannt, wobei schulartspezifische Unterschiede in der Testteilnahmemotivation in Low-Stakes-Tests diskutiert werden. Abschließend werden Strategien vorgestellt, wie mit unmotivierten Testteilnahmeverhalten in Low-Stakes-Assessments umgegangen werden kann.

### 7.4.1 Testteilnahmemotivation als Einschränkung valider Interpretationen der Testergebnisse in der Berichtserstattung?

Zusammenfassend kann auf Basis der Ergebnisse der drei Studien abgeleitet werden, dass Testteilnahmemotivation einen Zusammenhang mit der gezeigten Testleistung aufweist, auch wenn das Ausmaß des Einflusses der Testteilnahmemotivation zwischen den Studien variierte. Folglich macht Testteilnahmemotivation einen Teil konstrukt-irrelevanter Varianz in der Testleistung aus. Die Ergebnisse werden dadurch bestärkt, dass repräsentative Stichproben vorlagen und die Teilnehmenden an einer mehrstündigen Testung mit einer entsprechend hohen Aufgabenanzahl teilnahmen. Einschränkend muss allerdings erwähnt werden, dass in den Studien nicht für die Vorleistung kontrolliert werden konnte, da keine Ergebnisse aus standardisierten High-Stakes-Tests vorlagen. Um eine Annäherung an solch einen Wert zu bekommen, wurde in Studie III die Mathematiknote verwendet, da diese aus mehrerer High-Stakes-Tests aggregiert wird, auch wenn neben der Bewertung der Leistung unter anderem motivationale Faktoren in die Notengebung miteinfließen (Spinath, 2012). Auch wenn nicht für die Vorleistung in Form eines standardisierten High-Stakes-Testwertes kontrolliert werden konnte, wurde in der finalen Analyse in Studie III diverse Hintergrundvariablen berücksichtigt. Diese Analyse ergab, dass auch unter Berücksichtigung des sozioökonomischen Status, des Zuwanderungshintergrunds, des Geschlechts, der Mathematiknote und der Schulart die Testteilnahmemotivation einen

eigenen Beitrag zur Aufklärung der Leistungsvarianz erbringt. Die Zusammenhänge zwischen den motivationalen Konstrukten und Testleistung lagen dabei in derselben Größenordnung wie die Beziehung zwischen Testleistung und Zuwanderungshintergrund beziehungsweise Mathematiknote.

Zukünftige Studien sollten daher für Testteilnahmemotivation korrigieren, wie es auch in den *Standards for Educational and Psychological Testing* explizit empfohlen wird. In Standard 10.12 wird zum Beispiel darauf hingewiesen, dass bei der Auswertung von Assessments und der Interpretation der Ergebnisse für die Berichterstattung das Ausmaß von Faktoren berücksichtigt werden sollte, die konstrukt-irrelevante Varianz in den Testergebnissen ausmachen können. „An additional type of information that is relevant to the interpretation of test results in policy settings is the degree of motivation of the test takers. It is important to determine whether test takers regard the test experience seriously, particularly when individual scores are not reported to test takers or when the scores are not associated with consequences for the test takers. Decision criteria regarding whether to include scores from individuals with questionable motivation should be clearly documented“ (AERA, APA & NCME, 2014, S. 213). Um in der Berichtslegung deskriptive Statistiken unverzerrt von motivationalen Einflüssen darzulegen, könnten die Daten von den Schülerinnen und Schülern entfernt werden, die unplausible Angaben bezüglich ihrer Testteilnahmemotivation oder eine sehr niedrige Testteilnahmemotivation berichteten. Dieses Vorgehen entspricht der in Abschnitt 7.5 diskutierten Motivationsfilterung. In Analysen, in denen die Zusammenhänge zwischen Variablen und Leistung aufgezeigt werden sollen, könnte die Testteilnahmemotivation auch als Kovariate in die Analysen aufgenommen werden. Auf diese Weise kann für Unterschiede in der Leistung der Schülerinnen und Schüler, die auf die Testteilnahmemotivation zurückgehen, kontrolliert werden. Auf Basis dieser Leistungsdaten, die den Einfluss der Testteilnahmemotivation berücksichtigen, kann eine valide Interpretation der Ergebnisse in Schulleistungsstudien eher gewährleistet werden.

### 7.4.2 Zeitliche Erfassung der Testteilnahmemotivation

Aus testökonomischen Gründen ist die Erfassung der Testteilnahmemotivation zu mehreren Messzeitpunkten in den meisten Studien nicht umsetzbar. Wenn also nicht die motivationalen Prozesse während der Testsitzung näher beleuchtet werden sollen, muss in der Praxis meist ein Messzeitpunkt ausreichend sein. Soll der Zusammenhang zwischen Testteilnahmemotivation und Leistung untersucht werden, wird aus den Ergebnissen der

Studien II und III die praktische Empfehlung abgeleitet, die Testteilnahmemotivation nach dem Test zu erheben. Der Zusammenhang zwischen der Erwartungskomponente und Testleistung war nach dem Test stärker, was darauf hindeutet, dass die Schülerinnen und Schüler ihre Erfolgserwartungen nach der Bearbeitung aller Testaufgaben realistischster einschätzen als nach der Instruktion vor dem Test, in der sie lediglich einige Beispielaufgaben vorgestellt bekommen. Damit scheint es für die bisher so oft vernachlässigte Erwartungskomponente geeigneter, diese nach dem Test zu erheben, wenn mehrere Messzeitpunkte nicht realisierbar sind. Der Zusammenhang zwischen Testleistung und Anstrengungsbereitschaft sowie der Wertkomponente verändert sich nicht, wenn die Messzeitpunkte vor dem Test und nach dem Test verglichen wurden. Daher ist sowohl für die Forschung als auch für das Bildungsmonitoring empfehlenswert, Testteilnahmemotivation nach dem Test zu erfassen.

### 7.4.3 Verknüpfung von domänenspezifischen Kompetenzüberzeugungen, situationsspezifische Erfolgserwartungen und Aufgabenschwierigkeiten

Dass domänenspezifische Kompetenzüberzeugungen, wie zum Beispiel das Selbstkonzept in einem Fach, mit der Leistung in der entsprechenden Domäne zusammenhängen, findet Anschluss an die internationale und nationale Forschung (Chen et al., 2013; Jansen et al., 2013). Im Erwartung-Wert-Modell wird postuliert, dass domänenspezifische Kompetenzüberzeugungen auf die situationsspezifische Erwartungskomponente Einfluss nehmen. Das heißt beispielsweise, dass das Selbstkonzept in Mathematik Einfluss auf die Erfolgserwartungen nimmt, einen bestimmten Mathematiktest zu meistern. In der ersten Studie konnte im PISA 2000 Kontext bestätigt werden, dass auch nach Kontrolle domänenspezifischer Kompetenzüberzeugungen die Anstrengungsbereitschaft und Wertkomponente einen Teil der gezeigten Testleistung erklären. Daher war eine weitere Betrachtung des Zusammenhangs zwischen Testleistung und Testteilnahmemotivation inklusive der Erwartungskomponente erforderlich. Überdies zeigten Studien (Asseburg, 2011; Eccles & Wigfield, 2002), dass domänenspezifische Kompetenzüberzeugungen situationsspezifische Erfolgserwartungen erklären können. Studie III ergab außerdem, dass die situationsspezifischen Erfolgserwartungen sowie deren Veränderung während einer Testsitzung einen Zusammenhang mit Testleistung aufweisen, auch nach Kontrolle domänenspezifischer Kompetenzüberzeugungen.

Für die Assessmentpraxis bedeutet das, dass neben einem hohen Selbstkonzept auch die spezifische, während des Tests wahrgenommene Zuversicht der Schülerinnen und Schüler (d. h. Erfolgserwartung) für eine hohe Testleistung wichtig ist. Weiterführende Überlegungen, wie die Erfolgserwartungen praktisch beeinflusst werden können, münden bei der Schwierigkeit des Tests beziehungsweise der Aufgaben. Die Aufgaben sollten insgesamt nicht zu schwierig sein, beziehungsweise sollten sich während des Tests leichte und schwierige Aufgaben abwechseln. Asseburg und Frey (2013) fanden in ihrer Untersuchung auf Basis von PISA-2006-Daten, dass für etwa zwei Drittel der Stichprobe die durchschnittliche Aufgabenschwierigkeit die individuelle Fähigkeit überstieg und der Test de facto zu schwer für die Neuntklässlerinnen und Neuntklässler war. Für Teilnehmende, deren Fähigkeit deutlich unter der Aufgabenschwierigkeit lag, zeigte sich ein negativer Zusammenhang zwischen der Testleistung und Anstrengung sowie Langeweile. Wenn der Test die Teilnehmenden überfordert, kann sich dies demnach in fehlender Anstrengungsbereitschaft widerspiegeln sowie durch geringe Erfolgserwartungen indiziert werden. Als praktische Implikation wird eine mittlere Lösungswahrscheinlichkeit der Aufgaben von 70% empfohlen (Asseburg & Frey, 2013).

Auch computeradaptive Tests berücksichtigen die Passung der Aufgabenschwierigkeit mit der Fähigkeit der Testteilnehmenden, indem auf Grundlage der gegebenen Antwort auf eine Aufgabe die nächste Aufgabe ausgewählt wird (Frey & Seitz, 2011). Demnach ist in diesem Kontext ein stabiler Verlauf der Erfolgserwartungen während der Testsitzung wahrscheinlich. Studie III ergab einen durchschnittlichen stabilen Verlauf der Erfolgserwartungen während des Tests. Allerdings waren interindividuelle Unterschiede in dieser intraindividuellen Veränderung der Erfolgserwartungen vorhanden, so dass die Veränderung in den Erfolgserwartungen einen Zusammenhang mit der Mathematikleistung aufwies. Diese interindividuellen Unterschiede im intraindividuellen Verlauf sollten sich in der Anwendung von computeradaptiven Tests minimieren. Häusler und Sommer (2008) fanden in ihrer Studie, dass der vermehrte Einsatz motivierender, leichter Aufgaben die wahrgenommene Erfolgswahrscheinlichkeit der Testeilnehmenden erhöhen kann, ohne dabei mehr Zeit für die Testbearbeitung zu benötigen, da leichte Aufgaben weniger Zeit beanspruchen. Sie betonen auch, dass eine Lösungswahrscheinlichkeit von 50% wahrscheinlich zu niedrig ist, um die Testteilnahmemotivation während des Tests aufrechtzuerhalten. Die Ergebnisse ihrer Studie weisen eher darauf hin, dass während des Tests vermehrt leichte Aufgaben mit einer Lösungswahrscheinlichkeit von mehr als 50%

eingesetzt werden sollten, um hohe Erfolgserwartungen der Testteilnehmenden zu sichern (Häusler & Sommer, 2008).

### 7.4.4 Schulartunterschiede in der Testteilnahmemotivation

Eine Forschungsfrage der ersten Studie befasste sich mit Unterschieden in der Testteilnahmemotivation zwischen Teilnehmenden an Gymnasien und Teilnehmenden an nicht-gymnasialen Schularten. Es bestanden keine Schulartunterschiede für den Zusammenhang zwischen der situationsspezifischen Testteilnahmemotivation und Leistung. Die Vorhersage der Anstrengungsbereitschaft für die zwei untersuchten Gruppen ergab, dass es für Schülerinnen und Schüler an nicht-gymnasialen Schularten wichtiger ist, dass der Test attraktiv gestaltet ist und ihnen die Bearbeitung der Testaufgaben Freude bereitet. Damit scheint der Interessenaspekt der Wertkomponente des Erwartung-Wert-Anstrengung-Modells für Teilnehmende an nicht-gymnasialen Schularten bedeutsamer zu sein als für Gymnasiastinnen und Gymnasiasten. Möglicherweise sind Gymnasiastinnen und Gymnasiasten durch die Unterrichtskultur am Gymnasium es mehr gewohnt, sich über längere Zeit auf eine Aufgabe zu konzentrieren, so dass die Ablenkung während des Tests eher für Teilnehmende anderer Schularten relevant wird. Auch wird der Unterricht an Gymnasien störungsfreier durchgeführt als an den anderen Schularten (Brunner, 2006).

Dieser Befund kann insofern positiv gedeutet werden, als dass es bei der Konstruktion von Schulleistungstests möglich ist, solche testbezogenen Aspekte, wie die Attraktivität eines Tests zu berücksichtigten. Beispielsweise wurde gezeigt, dass visuell anregende Aufgaben mit Grafiken und mit wenig Leseaufwand oder einer kleinen Anzahl an Antwortoptionen sich positiv auf die Anstrengungsbereitschaft auswirken (Wise et al., 2009). Allerdings beeinflussen diese Adaptionen auch die Aufgabenschwierigkeit, das heißt sie machen die Aufgabe leichter (Leucht et al., 2012; Prenzel, Häußler, Rost & Senkbeil, 2002). In Schulleistungsstudien kann der Leseaufwand sowie die Anzahl an Antwortoptionen nicht für jede Aufgabe reduziert werden, da so alle Aufgaben leichter werden würden. Dies stünde dann mit dem Anspruch dieser Studien im Widerspruch, ein möglichst breites Spektrum an Kompetenzen mit den Aufgaben zu erfassen. Eine Überlegung könnte allerdings sein, ob bei der Konstruktion von Testaufgaben auch teilweise dann illustrative Grafiken eingesetzt werden, wenn diese nicht unbedingt für die Bearbeitung der Aufgabe benötigt werden und nur zur Gestaltung des Testheftes beitragen. Wobei hier jedoch Kosten und Nutzen vor allem in Hinblick auf die zusätzlich anfallenden Ausgaben für den Druck der (eventuell auch farbigen) Bilder abgewogen werden müssen.

### 7.4.5 Umgang mit unmotiviertem Verhalten in Low-Stakes-Tests

In der Literatur gibt es mittlerweile eine Reihe an Strategien, die in Low-Stakes-Tests angewendet werden können, um einer niedrigen Testteilnahmemotivation vorzubeugen beziehungsweise, wie mit ihr umgegangen werden kann. Auf diese Strategien wird im Folgenden eingegangen. Nach van Barneveld, Pharand, Ruberto und Haggarty (2013) kann eine niedrige Testteilnahmemotivation sich in verschiedenen Verhaltensweisen der Schülerinnen und Schüler manifestieren. Beispielsweise kann es passieren, dass unmotivierte Testteilnehmende bei Aufgaben mit geschlossenen Antwortformaten wahllos Muster im Testheft markieren oder auch in den Schülerfragebögen, obwohl negativ formulierte Fragen oder Aussagen vorhanden sind. Ein anderes Beispiel für unmotiviertes Verhalten betrifft das Raten der Lösung von Aufgaben oder die schlichte Nichtbeantwortung von Aufgaben (van Barneveld et al., 2013). Wise (2009) benennt vier bekannte Strategien, um mit niedriger Testteilnahmemotivation umzugehen, die im Folgenden kurz vorgestellt werden: a) die Einführung von Konsequenzen beziehungsweise die Umwandlung des Low-Stakes-Tests in einen High-Stakes-Test, b) die Bereitstellung von Feedback, c) die Bereitstellung von Anreizen und d) die Betonung der *academic citizenship* der Testteilnehmenden, also die gesellschaftliche Wichtigkeit einer Testteilnahme, auch wenn der Test keine Konsequenzen nach sich zieht.

Bezüglich Strategie a) ergaben Studien (Knekta & Eklöf, im Druck; Sundre & Kitsantas, 2004; Wolf & Smith, 1995), dass die Testteilnahmemotivation in High-Stakes-Test meist höher ist als in Low-Stakes-Tests (s. Abschnitt 2.4.3), allerdings nicht immer in einer höheren Testleistung für die High-Stakes-Gruppe mündet (Baumert & Demmrich, 2001; Cole & Osterlind, 2008). Darüber hinaus ist die Einführung von Konsequenzen für viele Schulleistungsstudien nicht erwünscht aufgrund einer möglichen Erhöhung der Testangst.

Untersuchungen, die Strategie b) umsetzten und Feedback ankündigten (Baumert & Demmrich, 2001; Wise, 2004), fanden keine erhöhte Anstrengungsbereitschaft oder Testleistung in der Feedback-Gruppe im Vergleich der Kontrollgruppe, die kein Feedback angekündigt bekamen. Ebenfalls zeigte die Umsetzung von Strategie c), die Bereitstellung von monetären Anreizen, unterschiedliche Wirksamkeit: eine Erhöhung der Testleistung ergab sich nur, wenn Schülerinnen und Schüler der achten Klasse an der Studie teilnahmen; bei Teilnehmenden der zwölften Klasse konnte keine erhöhte Testleistung im Vergleich zur Kontrollgruppe festgestellt werden (O'Neil, Abedi, Miyoshi & Mastergeor-

ge, 2005; O'Neil, Sugrue & Baker, 1995). Außerdem ist zu bedenken, dass sowohl individuelles Feedback und Geldanreize für Large-Scale-Assessments aufgrund der hohen Teilnehmerzahl und der damit verbundenen, enormen Kosten nicht praktikabel sind.

Strategie d) scheint im Vergleich zu den anderen Strategien am besten in Schulleistungsstudien umsetzbar, da sie am kostengünstigsten ist, wenig Zeit benötigt und keine negativen Auswirkungen, wie die potentielle Erhöhung von Testangst, mit sich bringt. Untersuchungen, die an die *academic citizenship* der Teilnehmenden appellierten, indem sie in der Instruktion die Wichtigkeit des Tests betonten, zeigten jedoch keine Erhöhung in der Testteilnahmemotivation oder Testleistung (Baumert & Demmrich, 2001; Kornhauser, Minahan, Siedlecki & Steedle, 2014). In der Studie von Lau, Swerdzewski, Jones, Anderson und Markle (2009) wurde eine erweiterte Strategie zur Verbesserung der Testteilnahmemotivation entwickelt. Die Testleiterinnen und Testleiter betonten nicht nur die Wichtigkeit und Nützlichkeit des Tests, sondern bedankten sich auch bei den Teilnehmenden für die Bearbeitung des Tests und ermunterten sie explizit zur Testteilnahme. Dabei war es auch erwünscht, auf einzelne Teilnehmenden individuell einzugehen. Diese Strategie führte dazu, dass die Teilnehmenden eine höhere Testteilnahmemotivation berichteten. Ob die Teilnehmenden, die eine erhöhte Testteilnahmemotivation berichteten, auch eine höhere Testleistung zeigten als die Kontrollgruppe, wurde nicht berichtet. Auch wenn diese Strategie vielversprechend erscheint, kann durch das individuelle Eingehen der Testleitung auf die Teilnehmenden nicht mehr das übliche, standardisierte Vorgehen in einer Testsitzung gesichert werden, das notwendig ist, damit in jeder Schule beziehungsweise jeder Klasse der Test unter denselben Bedingungen administriert wird (Lüdtke, Robitzsch, Trautwein, Kreuter & Ihme, 2007). Ob sich die oben beschriebene Strategie zur Erhöhung der Testteilnahmemotivation der Schülerinnen und Schüler in Low-Stakes-Tests auf Kosten eines etwas geringeren Standardisierungsgrad der Testdurchführung während der Testsitzung lohnt, sollte in weiteren Studien überprüft werden.

Da diese vier beschriebenen Strategien in großangelegten Schulleistungsstudien schwierig umzusetzen sind beziehungsweise die Effektivität zur Erhöhung der Testteilnahmemotivation in der Assessmentpraxis weiterhin fraglich ist, wurden in den letzten Jahren neue Strategien entwickelt, um die Gefahr invalider Interpretationen der Testergebnisse von Low-Stakes-Assessments zu minimieren: e) *motivation filtering*, f) *rapid-response filtering*, g) *effort-moderated item response theory model* und h) *effort-monitoring computer-based tests* (Wise, 2009). Im nächsten Abschnitt werden diese vier

Strategien der Vollständigkeit halber kurz beschrieben, auch wenn drei der vier neuen Strategien computerbasiertes Testen voraussetzen. Bei Tests, die am Computer bearbeitet werden, kann zum Beispiel die aufgebrachte Antwortzeit für eine Aufgabe als Indikator der Testteilnahmemotivation gewertet werden (Wise, 2006; Wise et al., 2009). Wenn die Aufgabe unrealistisch schnell beantwortet wurde, so dass die Antwortzeit nicht einmal für das Durchlesen der Aufgabenstellung ausreichend gewesen wäre, kann von Rateverhalten (oder auch *rapid-guessing behaviour*) gesprochen werden. Solch ein Rateverhalten kann Einfluss auf den Schwierigkeitsparameter der Aufgaben dahingehend nehmen, dass die Aufgabe durch Falschraten schwieriger geschätzt werden, als wenn die Antwort nicht geraten wurde (van Barneveld et al., 2013).

Die Motivationsfilterung ist die einzige Strategie (e), die auch in *Paper-and-Pencil-Tests* anwendbar ist. Dabei werden die gesamten Ergebnisse der Teilnehmenden aus den Daten entfernt (d. h. es wird der komplette Fall gelöscht), die geringe Anstrengungsbereitschaft berichten, was insgesamt zu einer höheren Leistung und zu einer Erhöhung der Validität der Testergebnisse führt (Kong, Wise & Bhola, 2007; Wise & DeMars, 2005; Wise & Kong, 2005; Wise, Wise & Bhola, 2006). Grundannahme für das Filtern der Daten ist, dass die Anstrengungsbereitschaft und die Fähigkeit der Teilnehmenden unkorreliert sind (Wise, 2009). Wenn diese Annahme verletzt ist und trotzdem die Daten gefiltert werden, verzerrt das Filtern die Verteilung der Fähigkeit „positiv“, da der Leistungsmittel nach oben verschoben wird. Swerdzewski et al. (2011) fanden, dass die selbstberichtete Anstrengungsbereitschaft der Studierenden, die mithilfe von Fragebögen erfasst wurde, auch ein gültiges Maß für ihre Motivation darstellte wie die durch den Computer aufgezeichnete Aufgabenantwortzeit. In ihrer Untersuchung kam die Motivationsfilterung auf Basis der Selbsteinschätzungen zu denselben Ergebnissen wie die Motivationsfilterung auf Basis der aufgezeichneten Antwortzeit. Möglicherweise könnte bei der Motivationsfilterung auf Grundlage der Anstrengungsbereitschaft die Erfolgswahrscheinlichkeit mit berücksichtigt werden. So können mehr Informationen über den motivationalen Zustand der Teilnehmenden gewonnen werden, da Nichtanstrengung nicht nur ein Zeichen von Langeweile sondern auch von sinkender Erfolgserwartung sein kann. So könnten sich die Teilnehmenden aufgrund eines zu schweren Tests nicht anstrengen, führt es möglicherweise nicht zwangsläufig zu einer Erhöhung der validen Interpretation der Ergebnisse, wenn diese Teilnehmenden aus den Daten entfernt werden. An dieser Stelle ist allerdings weitere Forschung notwendig.

Strategie f), das *rapid-response filtering* hat Ähnlichkeiten mit der Motivationsfilterung, nur das nicht vollständige Fälle aussortiert werden wie in Strategie e), sondern nur die einzelnen Antworten, in denen Rateverhalten aufgezeichnet wurde. Allerdings wird diese Methode nicht empfohlen, da zum einen nur ein schwacher Einfluss auf die konvergente Validität gefunden wurde (Kong, 2007) und zum anderen die Reliabilität der Personenparameter unklar ist, da pro Person die Schätzung der Fähigkeit auf unterschiedlichen Sets von Aufgaben basierten (Wise, 2009).

In dem von Wise und DeMars (2006) entwickelten *effort-moderated item response theory model* (EMIRT; Strategie g) wird die benötigte Antwortzeit in die Schätzung der Personenfähigkeit und der Aufgabenschwierigkeit miteinbezogen. Ihre Studie zeigte, dass, wenn in den Daten Rateverhalten modelliert wurde, das EMIRT einen besseren Modellfit zeigte sowie genauere Itemparameterschätzungen und eine höhere Validität als das Standard-IRT-Modell. Die Anwendung dieses vielversprechenden Modells setzt jedoch eine computerbasierte Testung voraus.

Die letzte Strategie (h), die kurz vorgestellt werden soll, betrifft *effort-monitoring computer-based tests*. Dies ist eine proaktive Methode, um Nichtanstrengungen während einer computerbasierten Testung zu reduzieren. Die von Wise, Bhola und Yang (2006) entwickelte Methode beinhaltet Warnmeldungen, die auf den Bildschirmen der Testteilnehmenden erscheinen, die Rateverhalten zeigen. Dabei führte das Einblenden einer Warnmitteilung dazu, dass weniger Rateverhalten während des Tests gezeigt wurde und die Validität der Interpretation der Ergebnisse dieser Gruppe stieg im Vergleich zur Kontrollgruppe, die keinerlei Meldungen bekam. Eine weitere Studie (Kong, Wise, Harmes & Yang, 2006) fand zusätzlich, dass die Teilnehmenden mit mindestens einer Warnmeldung auch eine höhere Testleistung zeigten als die Kontrollgruppe. Sie prüften darüber hinaus auch den Effekt von lobenden Hinweisen. Die Gruppe, die eine lobende Mitteilung bei durchgängigem Lösungsverhalten erhielt, berichtete keinen Anstieg in der Leistung oder der Anstrengungsbereitschaft im Vergleich zur Kontrollgruppe (Kong, Wise, Harmes & Yang, 2006). Der Vorteil dieser Strategie besteht darin, dass, im Gegensatz zum Filtern, im Vorfeld keine Annahme über den Zusammenhang von Anstrengungsbereitschaft und Fähigkeit getroffen werden muss. Das heißt, die Strategie h) kann immer angewendet werden, auch wenn in den Daten die für das Filtern notwendige Annahme der Nullkorrelation zwischen Anstrengungsbereitschaft und Fähigkeit verletzt ist.

Abschließend bleibt festzuhalten, dass computerbasierte Tests neue Perspektiven für die Erfassung und Beeinflussung von Testteilnahmemotivation eröffnen. Vor allem durch die unauffällige Aufzeichnung der Zeit, die Testteilnehmende für ihre Antwort benötigen, als Ersatz oder als ergänzendes Maß zur Selbstauskunft der Schülerinnen und Schüler über ihrer Testteilnahmemotivation können feinschrittige Informationen über die Motivation sowie über den Verlauf der Testteilnahmemotivation gewonnen werden. Außerdem erscheint die unaufwändige und im Vergleich zu anderen Strategien kostengünstige Beeinflussung der Anstrengungsbereitschaft durch den Einsatz von Warnmeldungen sehr vielversprechend. Allerdings sollte den Forscherinnen und Forschern bewusst sein, dass bei alleiniger Betrachtung der Antwortzeit als Maß der Anstrengungsbereitschaft nur ein Aspekt des vielschichtigen Konstrukts der Testteilnahmemotivation untersucht wird.

## 7.5 Ausblick und Fazit

Zum Abschluss dieser Arbeit werden weitere Forschungsfelder aufgezeigt, die zum einen direkt an die hier durchgeführten Studien anschließen können, wie eine Untersuchung der postulierten Interaktion von Erwartung und Wert. Zum anderen wird auf Anschlussmöglichkeiten der Testteilnahmemotivation an stabile Personeneigenschaften und an andere situationsspezifische Merkmale hingewiesen, bevor die Arbeit mit einem Fazit endet.

### 7.5.1 Erwartung-mal-Wert-Interaktion in der Testteilnahmemotivation

Das moderne Erwartung-Wert-Modell (Wigfield & Eccles, 2000) der Leistungsmotivation sowie das hier entworfene Erwartung-Wert-Anstrengung-Modell der Testteilnahmemotivation (s. Abschnitt 2.4.4) setzen die Gültigkeit eines additiven Modells voraus. Bei einem additiven Modell genügt es, dass entweder die Erwartungskomponente oder die Wertkomponente hoch ausgeprägt ist, damit eine hohe Motivation resultiert. Allerdings wurde ursprünglich im Risikowahl-Modell eine multiplikative Verknüpfung von Erwartung und Wert angenommen (Atkinson, 1957), so dass eine hohe Leistungsmotivation nur dann resultierte, wenn Erwartung und Wert hoch ausgeprägt sind (s. Abschnitt 2.3.1).

Mit dem Übergang zum modernen Erwartung-Wert-Modell (Eccles & Wigfield, 2002) wurde Mitte der 1980er Jahre diese Interaktion in eine additive Verbindung umgewandelt. Nagengast, Marsh, Xu, Hau und Trautwein (2011) versuchten, das „verlorengegangene mal“ in die Erwartung-mal-Wert-Theorie zurückzubringen. Unter Verwendung der Daten von 57 Ländern aus PISA 2006 zeigten ihre Analysen, dass die multiplikative Beziehung von Erwartung (Selbstkonzept in Naturwissenschaften) und Wert (Freude an

Naturwissenschaften) einen statistisch signifikant positiven, wenn auch kleinen Einfluss zeigte auf a) die außerschulischen Aktivitäten im Bereich Naturwissenschaften und b) auf die Absicht, eine naturwissenschaftliche Karriere anzugstreben. Das bedeutet, dass der Effekt des Selbstkonzepts in Naturwissenschaften auf a) und b) höher wurde, wenn die Freude an Naturwissenschaften ebenfalls hoch war sowie umgekehrt. Der multiplikative Zusammenhang trat vor allem für die außerschulischen Aktivitäten zutage: Wenn die Schülerinnen und Schüler wenig Freude an Naturwissenschaften berichteten, zeigte das Selbstkonzept in Naturwissenschaften keinerlei Effekt auf das Engagement; war die berichtete Freude hoch, ergab ein hohes Selbstkonzept auch eine erhöhte Bereitschaft zu außerschulischen Aktivitäten (Nagengast et al., 2011). Auch Trautwein et al. (2012) untersuchten den Einfluss der Interaktion von Selbstkonzept (Erwartung) und allen vier Wertaspekten (Wichtigkeit, Interesse, Nützlichkeit und Kosten) auf die Leistung in Mathematik und Englisch von Gymnasiastinnen und Gymnasiasten der Sekundarstufe II. Jeder der vier Wertaspekte zeigte eine Interaktion mit Selbstkonzept, die positiv mit der Testleistung zusammenhing, auch unter Kontrolle diverser Hintergrundvariablen (z. B. Vorleistung und schlussfolgerndes Denken). Damit findet der „verlorengegangene“ Interaktionseffekt empirische Unterstützung in Studien zur domänenspezifischen Leistungsmotivation.

Ob eine Interaktion von Erwartung und Wert im Kontext der Testteilnahmemotivation vorliegt, ist bislang unklar. Wenn das hier angepasste Erwartung-Wert-Anstrengung-Modell die Basis der Untersuchungen bildet, dann wäre zunächst zu prüfen, ob die Interaktionen zwischen Erwartung und den verschiedenen Wertaspekten einen signifikanten Zusammenhang mit der Anstrengungsbereitschaft und/oder mit der Testleistung zeigen.

### 7.5.2 Testteilnahmemotivation und Positionseffekte

Positionseffekte von Aufgaben gehören zu den Kontexteffekten (Brennan, 1992) und beziehen sich auf den Umstand, dass die Schwierigkeit einer Aufgabe in Abhängigkeit von ihrer Position im Testheft variiert. Dabei wird angenommen, dass Aufgaben am Ende eines Testheftes meist schwerer sind, als wenn sie am Anfang des Testheftes eingesetzt werden. Da im Raschmodell Aufgabenschwierigkeit und Personenfähigkeit auf einer gemeinsamen Skala abgebildet werden (Yen & Fitzpatrick, 2006), ist eine andere Sichtweise, dass die Aufgabenschwierigkeit während des Tests konstant ist, aber die Personen am Ende eines Testhefts die Aufgaben seltener lösen, beispielsweise aufgrund von Ermüdungseffekten

oder mangelnder Testteilnahmemotivation. Damit stellen Positionseffekte analog zu Testteilnahmemotivation eine potentielle Quelle konstrukt-irrelevanter Varianz dar.

Eine aktuelle Studie untersuchte auf Basis der Ländervergleichsdaten aus dem Jahr 2012, ob Positionseffekte durch die Veränderung in der Anstrengungsbereitschaft moderiert werden (Weirich, Penk, Hecht, Roppelt & Böhme, 2015). Die Ergebnisse zeigten, dass Positionseffekte der Aufgaben für solche Personen stärker ausgeprägt sind, die a) eine niedrige anfängliche Anstrengungsbereitschaft berichteten und b) eine größere Abnahme in der Anstrengungsbereitschaft zeigten als der Durchschnitt der Teilnehmenden. Schülerinnen und Schüler, die schon mit einer geringen Anstrengungsbereitschaft in den Test gehen und deren Anstrengungsbereitschaft im Verlauf abnimmt, lösen die Aufgaben im hinteren Teil des Testheftes tendenziell nicht, wodurch die Positionseffekte der Aufgaben stärker hervortreten. Dazu wären Untersuchungen wünschenswert, die neben der berichteten Anstrengung auch andere Komponenten des Erwartung-Wert-Anstrengung-Modells, wie zum Beispiel die anfängliche Erfolgserwartungen und deren Verlauf während der Testsitzung mit einbeziehen. Wie Studie III zeigte, gibt es nämlich keinen Zusammenhang mehr zwischen der Veränderung in der Anstrengungsbereitschaft und Testleistung, wenn auch die Veränderung in der Wichtigkeit des Tests und in der Erfolgserwartung mit in die Analyse einbezogen werden. Demnach ist es denkbar, dass die Veränderung in den Erfolgserwartungen den Einfluss der Veränderung in der Anstrengungsbereitschaft „auffängt“.

### 7.5.3 Testteilnahmemotivation und Persönlichkeitseigenschaften

Aktuelle Forschung im Kontext von Low-Stakes-Assessments versucht, die Testteilnahmemotivation der Teilnehmenden mit Persönlichkeitseigenschaften zu erklären. In der Persönlichkeitsforschung hat sich vor allem das Fünf-Faktoren-Modell durchgesetzt, das auch *Big Five* genannt wird (Digman, 1990). Nach diesem Modell konstituieren fünf Dimensionen die Struktur der Persönlichkeit: Verträglichkeit, Gewissenhaftigkeit, Offenheit, Extraversion und Neurotizismus. Im Kontext der Teilnahme an Low-Stakes-Test könnte erwartet werden, dass Testteilnehmende, die verträglich und gewissenhaft sind, eher Aufgaben eines Tests ohne persönliche Konsequenzen beantworten (DeMars, Bashkov & Socha, 2013). Diese Annahme wurde von Barry und Finney (im Druck) bestätigt, deren Untersuchung ergab, dass Studierende mir hoher Ausprägung auf der Verträglichkeits- und Gewissenhaftigkeitsskala des *Big Five Inventory* (John & Srivastava, 1999) auch zu einer hohen Anstrengungsbereitschaft tendierten. Ebenfalls zeigten die

Studierenden, die eine hohe Verträglichkeit berichteten, einen positiveren Verlauf der Anstrengungsbereitschaft (d. h. einen Anstieg) während der Testsitzung als weniger verträgliche Studierende (Barry & Finney, im Druck).

Die Persönlichkeitseigenschaften von Testteilnehmenden werden überdies verwendet, um Geschlechterunterschiede in der Testteilnahmemotivation zu erklären. Bisherige Studien zeigten, dass männliche Teilnehmende meist weniger Anstrengungsbereitschaft berichteten als weibliche Teilnehmende (Butler & Adams, 2007; DeMars et al., 2013) und dass der Zusammenhang zwischen Anstrengungsbereitschaft und Leistung für männliche Teilnehmende höher ist als für weibliche Teilnehmende (Eklöf, 2007; Eklöf & Nyroos, 2013). Die Untersuchung von DeMars, Bashkov und Socha (2013) ergab, dass Persönlichkeitseigenschaften wenigstens teilweise erklären, warum weibliche Teilnehmende mehr Anstrengungsbereitschaft in Low-Stakes-Assessments investieren als männliche Teilnehmende: Weibliche Studierende berichteten meist eine höhere Gewissenhaftigkeit und Verträglichkeit sowie weniger Arbeitsvermeidung als männliche Studierende. Die Ergebnisse zeigten zudem eine positive, wenn auch geringe Beziehung zwischen Gewissenhaftigkeit beziehungsweise Verträglichkeit und Anstrengung sowie eine negative Beziehung zwischen Arbeitsvermeidung und Anstrengung. Interessant für weitere Forschung zu Geschlechterunterschieden in Testteilnahmemotivation ist, ob Persönlichkeitseigenschaften neben den Unterschieden in der Anstrengungsbereitschaft auch Unterschiede in den Erfolgserwartungen oder den verschiedenen Wertaspekten erklären können. So können tiefere Einblicke in die motivationalen Charakteristika der Teilnehmenden durch die gleichzeitige Berücksichtigung von stabilen Personenmerkmalen und veränderlichen Situationsmerkmalen aufgedeckt werden, wie im Grundmodell der Motivationspsychologie vorgesehen.

Neben der identifizierten Verbindung von Persönlichkeitsmerkmalen und Anstrengungsbereitschaft, zeigten diese Studien auch eine Verknüpfung von Testteilnahmemotivation mit verschiedenen Zielen (Lernziele und Arbeitsvermeidungsziele). Ob Ziele auch in Low-Stakes-Assessments für Schülerinnen und Schüler der Sekundarstufe I eine Verbindung mit Erwartung, Wert und Anstrengung aufweisen, ist ein weiteres Forschungsfeld, das Potential für vertiefte Analysen birgt. In dem originalen Erwartung-Wert-Modell von Eccles und Wigfield (2002) werden Ziele explizit als Teil der motivationalen Überzeugungen berücksichtigt, die einen Einfluss auf die Erwartungskomponente aufweisen, so dass

auch hier eine gemeinsame Betrachtung von Personen- und Situationsmerkmalen Aufschlüsse über die Motivation der Testteilnehmenden bringen kann.

### 7.5.4 Testteilnahmemotivation und Emotionen

Neben den stabilen Personeneigenschaften, die einen Zusammenhang mit Testteilnahmemotivation aufweisen, können auch situative Merkmale die Motivation während der Testbearbeitung beeinflussen, wie zum Beispiel Emotionen, die während der Leistungssituation auftreten. Leistungsemotionen werden definiert als Emotionen, die sich auf eine kompetenzbezogene Tätigkeit oder ein Ergebnis beziehen (Pekrun, 2006). Zwar ergab Studie I, dass der emotionale Zustand weder mit Leistung noch mit Anstrengungsbereitschaft zusammenhing, aber es wurde lediglich eine Emotion, nämlich die Freude beziehungsweise der Optimismus der Teilnehmenden während der Testbearbeitung erfasst. Die Untersuchung von Pekrun, Elliot und Maier (2009) zeigte beispielsweise, dass unterschiedliche Leistungsemotionen, wie Freude, Langeweile, Ärger, Hoffnung, Stolz oder Angst, Testleistung vorhersagen. Die Erhebung verschiedener Leistungsemotionen während einer Low-Stakes-Testsitzung könnte in weiterführenden Analysen verwendet werden, um die investierte Testteilnahmemotivation näher zu erforschen. So ist es wahrscheinlich, dass neben der Erwartung- und Wertkomponente auch Emotionen, die während der Bearbeitung eines zweistündigen Leistungstests einsetzen und sich möglicherweise verändern, die investierte Anstrengungsbereitschaft und damit vielleicht auch die Leistung beeinflussen können. Es wäre auch möglich, dass die Erfolgserwartungen, die in dieser Arbeit einen starken Zusammenhang mit Leistung zeigten, auch durch diverse Leistungsemotionen vorhergesagt werden können.

### 7.5.5 Fazit

Die vorliegende Arbeit hatte die Untersuchung des Einflusses von Testteilnahmemotivation auf die tatsächlich gezeigte Testleistung in großangelegten Schulleistungsstudien zum Ziel. Den Ausgangspunkt bildete der vermehrte Einsatz nationaler und internationaler Low-Stakes-Assessments in deutschen Schulen seit der Jahrtausendwende und die damit verbundenen potentielle Gefahr, dass Testteilnahmemotivation als Teil konstruktirrelevanter Varianz die valide Interpretation und Nutzung der Testergebnisse einschränkt. Studien in diesem Bereich basieren meist auf dem Erwartung-Wert-Modell der Leistungsmotivation, jedoch mangelte es an einem theoretischen Modell, das die Besonderheiten der Testteilnahmemotivation berücksichtigt. Daher wurde in der vorliegenden Arbeit basierend

auf der Erwartung-Wert-Theorie ein Erwartung-Wert-Anstrengung-Modell der Testteilnahmemotivation entworfen, das in drei Studien empirisch untersucht wurde. In zwei authentischen Testsituationen wurde dabei sowohl die Erwartungs- und Wertkomponente als auch die Anstrengungsbereitschaft der Schülerinnen und Schüler erfragt und diese motivationalen Konstrukte miteinander sowie mit der tatsächlich gezeigten Testleistung in Verbindung gesetzt.

Insgesamt konnten die Ergebnisse der drei empirischen Studien das theoretisch postulierte Erwartung-Wert-Anstrengung-Modell fast vollständig bestätigen. Dabei umfasst der theoretische Mehrwert der Arbeit vor allem die Erkenntnis, dass Untersuchungen, die auf der Erwartung-Wert-Theorie beruhen, die Erwartungskomponente neben der Wertkomponente und der Anstrengungsbereitschaft berücksichtigen sollten. Der formulierte Einwand, dass die Erfolgserwartungen in Low-Stakes-Tests nicht ausschlaggebend sind, da die Teilnehmenden keine Rückmeldung zu ihrer Leistung erhalten und sie somit ihre Erfolgserwartungen nicht einschätzen können, konnte nicht bestätigt werden. Darüber hinaus zeigte sich ein Einfluss der Veränderungen der Erfolgserwartungen während der Testsitzung auf die gezeigte Testleistung, was erneut die Wichtigkeit dieser Komponente untermauert. Um demzufolge das komplexe Beziehungsgefüge zwischen Erwartung, Wert und Anstrengung sowie deren Beziehung mit der Testleistung komplett erfassen können, sind alle drei Aspekte der Testteilnahmemotivation bedeutsam. Das für diese Arbeit aufgestellte Erwartung-Wert-Anstrengung-Modell der Testteilnahmemotivation sollte auch für zukünftige Untersuchungen als theoretische Grundlage verwendet und erweitert werden.

Zusammenfassend lässt sich auf Grundlage der drei Studien schlussfolgern, dass Testteilnahmemotivation einen Zusammenhang mit der gezeigten Testleistung aufweist, auch nach Kontrolle des sozioökonomischen Hintergrundes, der domänenspezifischen Kompetenzüberzeugungen sowie der domänenspezifischen Leistung der Schülerinnen und Schüler. Demnach kann eine Gefährdung der validen Interpretation und Nutzung der Testergebnisse aufgrund konstrukt-irrelevanter Varianz in Form von Testteilnahmemotivation nicht ausgeschlossen werden. In zukünftigen Low-Stakes-Assessments sollte Testteilnahmemotivation unbedingt erhoben und in den Analysen zum Kompetenzstand der Schülerinnen und Schüler für die Berichtserstattung als Kontroll- oder Filtervariable einbezogen werden.

# Literatur

Abdelfattah, F. (2010). The relationship between motivation and achievement in low-stakes examinations. *Social Behavior and Personality: An International Journal*, *38*(2), 159–167. doi:10.2224/sbp.2010.38.2.159

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: An empirical examination of effects on performance and test-taker reactions. *Journal of Experimental Psychology: Applied*, *15*(2), 163–181. doi:10.1037/a0015719

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington, DC: AERA.

Asseburg, R. (2011). *Motivation zur Testbearbeitung in adaptiven und nicht-adaptiven Leistungstests* (Doctoral dissertation). Christian-Albrechts-Universität zu Kiel. Retrieved from the website http://d-nb.info/1013153863/34

Asseburg, R., & Frey, A. (2013). Too hard, too easy, or just right? The relationship between effort or boredom and ability-difficulty fit. *Psychological Test and Assessment Modeling*, *55*(1), 92–104.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, *64*(6), 359–372.

Atkinson, J. W. (1964). *An introduction to motivation*. New York: Van Nostrand.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. doi:10.1037/0033-295X.84.2.191

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: W.H. Freeman.

Barry, C. L. (2010). *Examining change in motivation across the course of a low-stakes testing session: An application of latent growth modeling* (Unpublished doctoral dissertation). James Madison University.

Barry, C. L., & Finney, S. J. (in press). Modeling change in effort across a low-stakes testing session: A latent growth curve modeling approach. *Applied Measurement in Education*.

Barry, C. L., Horst, S. J., Finney, S. J., Brown, A. R., & Kopp, J. P. (2010). Do examinees have similar test-taking effort? A high-stakes question for low-stakes testing. *International Journal of Testing*, *10*(4), 342–363. doi:10.1080/15305058.2010.508569

Baumert, J., & Demmrich, A. (2001). Test motivation in the assessment of student skills: The effects of incentives on motivation and performance. *European Journal of Psychology of Education*, *16*(3), 441–462.

Beckmann, J., & Heckhausen, H. (2006). Motivation durch Erwartung und Anreiz. In *Motivation und Handeln* (3rd ed., pp. 105–142). Heidelberg: Springer.

Böhm-Kasper, O., & Weishaupt, H. (2008). Quantitative Ansätze und Methoden in der Schulforschung. In W. Helsper & J. Böhme (Eds.), *Handbuch der Schulforschung* (2nd ed., pp. 91–123). Wiesbaden: VS Verlag für Sozialwissenschaften.

Bong, M., & Skaalvik, E. M. (2003). Academic self-concept and self-efficacy: How different are they really? *Educational Psychology Review*, *15*(1), 1–40.

Brennan, R. (1992). The Context of Context Effects. *Applied Measurement in Education*, *5*(3), 225–264. doi:10.1207/s15324818ame0503_4

Brunner, M. (2006). *Mathematische Schülerleistung: Struktur, Schulformunterschiede und Validität* (Doctoral dissertation). Humboldt-Universität zu Berlin, Berlin. Retrieved from the website http://library.mpib-berlin.mpg.de/ft/mbr/MBR_Mathematische_2006.pdf

Brunstein, J. C., & Heckhausen, H. (2006). Leistungsmotivation. In *Motivation und Handeln* (3rd ed., pp. 143–191). Heidelberg: Springer.

Butler, J., & Adams, R. J. (2007). The impact of differential investment of student effort on the outcomes of international studies. *Journal of Applied Measurement*, *8*(3), 279–304.

Cao, J., & Stokes, S. L. (2008). Bayesian IRT guessing models for partial guessing behaviors. *Psychometrika*, *73*(2), 209–230. doi:10.1007/s11336-007-9045-9

Chen, S.-K., Yeh, Y.-C., Hwang, F.-M., & Lin, S. S. J. (2013). The relationship between academic self-concept and achievement: A multicohort–multioccasion study. *Learning and Individual Differences*, *23*, 172–178. doi:10.1016/j.lindif.2012.07.021

Cole, J. S., Bergin, D. A., & Whittaker, T. A. (2008). Predicting student achievement for low stakes tests with effort and task value. *Contemporary Educational Psychology*, *33*(4), 609–624.

Cole, J. S., & Osterlind, S. J. (2008). Investigating differences between low- and high-stakes test performance on a general education exam. *The Journal of General Education*, *57*(2), 119–130. doi:10.1353/jge.0.0018

Debeer, D., & Janssen, R. (2013). Modeling item-position effects within an IRT framework. *Journal of Educational Measurement*, *50*(2), 164–185. doi:10.1111/jedm.12009

DeMars, C. E., Bashkov, B. M., & Socha, A. B. (2013). The role of gender in test-taking motivation under low-stakes conditions. *Research & Practice in Assessment*, *8*, 69–82.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, *41*(1), 417–440. doi:10.1146/annurev.ps.41.020190.002221

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, *41*(10), 1040–1048. doi:10.1037/0003-066X.41.10.1040

Eccles, J. S. (2005). Subjective task value and the Eccles et al. model of achievement-related choices. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 105–121). New York: Guilford Press.

Eccles, J. S., & Wigfield, A. (1995). In the mind of the actor: The structure of adolescents' achievement task values and expectancy-related beliefs. *Personality and Social Psychology Bulletin*, *21*(3), 215–225. doi:10.1177/0146167295213003

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109–132. doi:10.1146/annurev.psych.53.100901.135153

Eklöf, H. (2007). Test-taking motivation and mathematics performance in TIMSS 2003. *International Journal of Testing*, *7*(3), 311–326.

Eklöf, H. (2008). Test-taking motivation on low-stakes tests: A Swedish TIMSS 2003 example. In *Issues and methodologies in large-scale assessments, IERI Monograph Series* (Vol. 1, pp. 9–21). Hamburg: IEA-ETS Research Institute.

Eklöf, H. (2010a). Skill and will: Test-taking motivation and assessment quality. *Assessment in Education: Principles, Policy & Practice*, *17*(4), 345–356. doi:10.1080/0969594X.2010.516569

Eklöf, H. (2010b). *Student motivation and effort in the Swedish TIMSS Advanced field study*. Paper presented at the 4th IEA International Research Conference, Gothenburg.

Eklöf, H., & Nyroos, M. (2013). Pupil perceptions of national tests in science: Perceived importance, invested effort, and test anxiety. *European Journal of Psychology of Education*, *28*(2), 497–510. doi:10.1007/s10212-012-0125-6

Eklöf, H., Pavešič, B. J., & Grønmo, L. S. (2014). A cross-national comparison of reported effort and mathematics performance in TIMSS advanced. *Applied Measurement in Education*, *27*(1), 31–45. doi:10.1080/08957347.2013.853070

Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist*, *34*(3), 169–189. doi:10.1207/s15326985ep3403_3

Elliot, A. J., & Harackiewicz, J. M. (1996). Approach and avoidance achievement goals and intrinsic motivation: A mediational analysis. *Journal of Personality and Social Psychology*, *70*(3), 461–475. doi:10.1037/0022-3514.70.3.461

Erwin, T. D., & Wise, S. L. (2002). A Scholar-Practitioner Model for Assessment. In T. W. Banta (Eds.), *Building a scholarship of assessment* (pp. 67–81). San Francisco: Jossey-Bass.

Freund, P. A., & Holling, H. (2011). Who wants to take an intelligence test? Personality and achievement motivation in the context of ability testing. *Personality and Individual Differences*, *50*(5), 723–728. doi:10.1016/j.paid.2010.12.025

Freund, P. A., Kuhn, J. T., & Holling, H. (2011). Measuring current achievement motivation with the QCM: Short form development and investigation of measurement invariance. *Personality and Individual Differences*, *51*(5), 629–634. doi:10.1016/j.paid.2011.05.033

Frey, A., & Seitz, N.-N. (2011). Hypothetical use of multidimensional adaptive testing for the assessment of student achievement in the Programme for International Student Assessment. *Educational and Psychological Measurement*, *71*(3), 503–522. doi:10.1177/0013164410381521

Geiser, C., Keller, B. T., & Lockhart, G. (2013). First- versus second-order latent growth curve models: Some insights from latent state-trait theory. *Structural Equation Modeling: A Multidisciplinary Journal*, *20*(3), 479–503. doi:10.1080/10705511.2013.797832

Haag, L., & Götz, T. (2012). Mathe ist schwierig und Deutsch aktuell: Vergleichende Studie zur Charakterisierung von Schulfächern aus Schülersicht. *Psychologie in Erziehung und Unterricht*, *59*(1), 32–46. doi:10.2378/peu2012.art03d

Harackiewicz, J. M., Durik, A. M., Barron, K. E., Linnenbrink-Garcia, L., & Tauer, J. M. (2008). The role of achievement goals in the development of interest: Reciprocal relations between achievement goals, interest, and performance. *Journal of Educational Psychology*, *100*(1), 105–122. doi:10.1037/0022-0663.100.1.105

Häusler, J., & Sommer, M. (2008). The effect of success probability on test economy and self-confidence in computerized adaptive tests. *Psychology Science Quarterly*, *50*(1), 75–87.

Heckhausen, J., & Heckhausen, H. (2006). Motivation und Handeln: Einführung und Überblick. In J. Heckhausen & H. Heckhausen (Eds.), *Motivation und Handeln* (3rd ed., pp. 1–9). Heidelberg: Springer.

Hidi, S., & Harackiewicz, J. M. (2000). Motivating the academically unmotivated: A critical issue for the 21st century. *Review of Educational Research*, *70*(2), 151–179. doi:10.3102/00346543070002151

Hodapp, V., Glanzmann, P. G., & Laux, L. (1995). Theory and measurement of test anxiety as a situation-specific trait. In C. D. Spielberger & P. R. Vagg (Eds.), *Test anxiety: theory, assessment, and treatment* (pp. 47–58). Washington, DC: Taylor & Francis.

Hohensinn, C., Kubinger, K. D., Reif, M., Schleicher, E., & Khorramdel, L. (2011). Analysing item position effects due to test booklet design within large-scale assessment. *Educational Research and Evaluation*, *17*(6), 497–509. doi:10.1080/13803611.2011.632668

Horst, S. J. (2010). *A mixture-modeling approach to exploring test-taking motivation in large-scale low-stakes contexts* (Unpublished doctoral dissertation). James Madison University, Harrisonburg.

International Test Commission. (2001). International guidelines for test use. *International Journal of Testing*, *1*(2), 93–114. doi:10.1207/S15327574IJT0102_1

Jansen, M., Schroeders, U., & Stanat, P. (2013). Motivationale Schülermerkmale in Mathematik und den Naturwissenschaften. In H. A. Pant, P. Stanat, U. Schroeders, A. Roppelt, T. Siegle, & C. Pöhlmann (Eds.), *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I* (pp. 347–365). Münster: Waxmann.

John, O. P., & Srivastava, S. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin & O. P. John (Eds.), *Handbook of personality: Theory and research*. New York: Guilford Press.

Klieme, E., Baumert, J., Köller, O., & Bos, W. (2000). Mathematische und naturwissenschaftliche Grundbildung: Konzeptuelle Grundlagen und die Erfassung und Skalierung von Kompetenzen. In J. Baumert, W. Bos, & R. Lehmann (Eds.), *Dritte Internationale Mathematik- und Naturwissenschaftsstudie – Mathematische und naturwissenschaftliche Grundbildung am Ende der Schullaufbahn: Bd. 1. Mathematische und naturwissenschaftliche Grundbildung am Ende der Pflichtschulzeit* (pp. 85–133). Opladen: Leske + Budrich.

KMK (2004) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005). *Bildungsstandards im Fach Mathematik für den Mittleren Schulabschluss. Beschluss vom 4.12.2003*. Luchterhand. Retrieved from http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Biologie.pdf

KMK (2005a) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005). *Bildungsstandards im Fach Mathematik für den Hauptschulabschluss. Beschluss vom 15.10.2004*. Luchterhand. Retrieved from http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Biologie.pdf

KMK (2005b) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005). *Bildungsstandards im Fach Biologie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004*. Luchterhand. Retrieved from http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Biologie.pdf

KMK (2005c) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005). *Bildungsstandards im Fach Chemie für den Mittleren Schulabschluss. Beschluss vom 16.12.2004*. Luchterhand. Retrieved from http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Biologie.pdf

KMK (2005d) = Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss. Beschluss vom 16.12.2004*. Luchterhand. Retrieved from http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2004/2004_12_16-Bildungsstandards-Biologie.pdf

Knekta, E., & Eklöf, H. (in press). Modeling the test-taking motivation construct through investigation of psychometric properties of an expectancy-value-based questionnaire. *Journal of Psychoeducational Assessment*.

Köller, O., & Baumert, J. (2012). Schulische Leistung und ihre Messung. In W. Schneider & U. Lindenberger (Eds.), *Entwicklungspsychologie* (Vol. 7, pp. 645–661). Weinheim: Beltz/PVU.

Köller, O., Baumert, J., & Schnabel, K. (2001). Does interest matter? The relationship between academic interest and achievement in mathematics. *Journal for Research in Mathematics Education*, *32*(5), 448. doi:10.2307/749801

Köller, O., Trautwein, U., Lüdke, O., & Baumert, J. (2006). Zum Zusammenspiel von schulischer Leistung, Selbstkonzept und Interesse in der gymnasialen Oberstufe. *Zeitschrift für Pädagogische Psychologie*, *20*(1/2), 27–39.

Kong, X. J. (2007). *Using response time and the effort-moderated model to investigate the effects of rapid guessing on estimation of item and person parameters* (Unpublished doctoral dissertation). James Madison University.

Kong, X. J., Wise, S. L., & Bhola, D. S. (2007). Setting the response time threshold parameter to differentiate solution behavior from rapid-guessing behavior. *Educational and Psychological Measurement*, *67*(4), 606–619. doi:10.1177/0013164406294779

Kong, X. J., Wise, S. L., Harmes, J. C., & Yang, S. (2006). *Motivational effects of praise in response-time-based feedback: A follow-up study of the effort-monitoring CBT*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Kornhauser, Z. G. C., Minahan, J., Siedlecki, K. L., & Steedle, J. T. (2014). *A strategy for increasing student motivation on low-stakes assessments*. Paper presented at the annual meeting of the American Educational Research Association, Philadelphia.

Lau, A. R., Swerdzewski, P. J., Jones, A. T., Anderson, R. D., & Markle, R. E. (2009). Proctors matter: Strategies for increasing examinee effort on general education program assessments. *The Journal of General Education*, *58*(3), 196–217. doi:10.1353/jge.0.0045

Leucht, M., Harsch, C., Pant, H. A., & Köller, O. (2012). Steuerung zukünftiger Aufgabenentwicklung durch Vorhersage der Schwierigkeiten eines Tests für die erste Fremdsprache Englisch durch Dutch Grid Merkmale. *Diagnostica*, *58*(1), 31–44. doi:10.1026/0012-1924/a000063

Linn, R. L. (2010). Validity. In P. Peterson, E. L. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (pp. 181–185). Oxford: Elsevier. Abgerufen von http://www.sciencedirect.com/science/referenceworks/9780080448947

Lüdtke, O., Robitzsch, A., Trautwein, U., Kreuter, F., & Ihme, J. M. (2007). Are there test administrator effects in large-scale educational assessments? *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *3*(4), 149–159. doi:10.1027/1614-2241.3.4.149

Marsh, H. W. (1990). The structure of academic self-concept: The Marsh/Shavelson model. *Journal of Educational Psychology*, *82*(4), 623–636. doi:10.1037/0022-0663.82.4.623

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child Development*, *76*(2), 397–416. doi:10.1111/j.1467-8624.2005.00853.x

Marsh, H. W., & Yeung, A. S. (1998). Longitudinal structural equation models of academic self-concept and achievement: Gender differences in the development of Math and English constructs. *American Educational Research Journal*, *35*(4), 705–738. doi:10.3102/00028312035004705

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5–11. doi:10.3102/0013189X018002005

Möller, J., & Trautwein, U. (2009). Selbstkonzept. In E. Wild & J. Möller (Eds.), *Pädagogische Psychologie* (pp. 170–203). Berlin: Springer.

Nagengast, B., Marsh, H. W., Scalas, L. F., Xu, M. K., Hau, K.-T., & Trautwein, U. (2011). Who took the „x“ out of expectancy-value theory? A psychological mystery, a substantive-methodological synergy, and a cross-national generalization. *Psychological Science*, *22*(8), 1058–1066. doi:10.1177/0956797611415540

Nicholls, J. G. (1984). Achievement motivation: Conceptions of ability, subjective experience, task choice, and performance. *Psychological Review*, *91*(3), 328–346. doi:10.1037/0033-295X.91.3.328

Nie, Y., Lau, S., & Liau, A. K. (2011). Role of academic self-efficacy in moderating the relation between task importance and test anxiety. *Learning and Individual Differences*, *21*(6), 736–741. doi:10.1016/j.lindif.2011.09.005

O’Neil, H. F., Abedi, J., Miyoshi, J., & Mastergeorge, A. (2005). Monetary incentives for low-stakes tests. *Educational Assessment*, *10*(3), 185–208. doi:10.1207/s15326977ea1003_3

O’Neil, H. F., Sugrue, B., & Baker, E. L. (1995). Effects of motivational interventions on the National Assessment of Educational Progress mathematics performance. *Educational Assessment*, *3*(2), 135–157.

Pajares, F., & Graham, L. (1999). Self-efficacy, motivation constructs, and mathematics performance of entering middle school students. *Contemporary Educational Psychology*, *24*(2), 124–139. doi:10.1006/ceps.1998.0991

Pajares, F., & Kranzler, J. (1995). Self-efficacy beliefs and general mental ability in mathematical problem-solving. *Contemporary Educational Psychology*, *20*(4), 426–443. doi:10.1006/ceps.1995.1029

Pant, H. A., Stanat, P., Schroeders, U., Roppelt, A., Siegle, T., & Pöhlmann, C. (2013). *IQB-Ländervergleich 2012. Mathematische und naturwissenschaftliche Kompetenzen am Ende der Sekundarstufe I.* Münster: Waxmann.

Pekrun, R. (2006). The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice. *Educational Psychology Review*, *18*(4), 315–341. doi:10.1007/s10648-006-9029-9

Pekrun, R., Elliot, A. J., & Maier, M. A. (2009). Achievement goals and achievement emotions: Testing a model of their joint relations with academic performance. *Journal of Educational Psychology*, *101*(1), 115–135. doi:10.1037/a0013383

Pintrich, P. R., Marx, R. W., & Boyle, R. A. (1993). Beyond cold conceptual change: The role of motivational beliefs and classroom contextual factors in the process of conceptual change. *Review of Educational Research*, *63*(2), 167–199. doi:10.3102/00346543063002167

Prenzel, M., Häußler, P., Rost, J., & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft*, *30*(2), 120–135.

Putwain, D. W., & Daniels, R. A. (2010). Is the relationship between competence beliefs and test anxiety influenced by goal orientation? *Learning and Individual Differences*, *20*(1), 8–13. doi:10.1016/j.lindif.2009.10.006

Rheinberg, F. (2008). *Motivation* (7th ed.). Stuttgart: W. Kohlhammer.

Rheinberg, F., Vollmeyer, R., & Burns, B. D. (2001). FAM: Ein Fragebogen zur Erfassung aktueller Motivation in Lern-und Leistungssituationen. *Diagnostica*, *47*(2), 57–66.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. doi:10.1037/0003-066X.55.1.68

Sayer, A. G., & Cumsille, P. E. (2001). Second-order latent growth models. In L. M. Collins & A. G. Sayer (Eds..), *New methods for the analysis of change* (1st ed., pp. 179–200). Washington, DC: American Psychological Association.

Schiefele, U. (1999). Interest and Learning From Text. *Scientific Studies of Reading*, *3*(3), 257–279. doi:10.1207/s1532799xssr0303_4

Schiefele, U. (2009). Motivation. In E. Wild & J. Möller (Eds.), *Pädagogische Psychologie* (pp. 151–177). Heidelberg: Springer.

Schunk, D. H., & Pajares, F. (2005). Competence perceptions and academic functioning. In A. J. Elliot & C. S. Dweck (Eds.), *Handbook of competence and motivation* (pp. 85–104). New York: Guilford Press.

Schunk, D. H., Pintrich, P. R., & Meece, J. L. (2008). *Motivation in education: Theory, research, and applications* (3rd ed.). Upper Saddle River, NJ: Pearson Education.

Schwippert, K., & Goy, M. (2008). Leistungsvergleichs- und Schulqualitätsforschung. In W. Helsper & J. Böhme (Eds.), *Handbuch der Schulforschung* (2nd ed., pp. 387–421). Wiesbaden: VS Verlag für Sozialwissenschaften.

Spinath, B. (2012). Academic achievement. In V. S. Ramachandran (Eds.), *Encyclopedia of human behavior* (pp. 1–8). London; Burlington, MA: Elsevier/Academic Press.

Stanat, P., & Lüdtke, O. (2013). International large-scale assessment studies of student achievement. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 481–483). New York, NY: Routledge.

Stiensmeier-Pelster, J., & Schöne, C. (2008). Fähigkeitsselbstkonzept. In W. Schneider & M. Hasselhorn (Eds.), *Pädagogische Psychologie* (pp. 50 – 73). Göttingen: Hogrefe-Verlag.

Sundre, D. L. (2007). *The Student Opinion Scale: A measure of examinee motivation: Test manual*. Retrieved from the Center for Assessment and Research Studies website: http://www.jmu.edu/assessment/resources/resource_files/sos_manual.pdf

Sundre, D. L., & Kitsantas, A. (2004). An exploration of the psychology of the examinee: Can examinee self-regulation and test-taking motivation predict consequential and non-consequential test performance? *Contemporary Educational Psychology*, *29*(1), 6–26. doi:10.1016/S0361-476X(02)00063-2

Swerdzewski, P. J., Harmes, J. C., & Finney, S. J. (2011). Two approaches for identifying low-motivated students in a low-stakes assessment context. *Applied Measurement in Education*, *24*(2), 162–188. doi:10.1080/08957347.2011.555217

Thelk, A. D., Sundre, D. L., Horst, S. J., & Finney, S. J. (2009). Motivation matters: Using the Student Opinion Scale to make valid inferences about student performance. *The Journal of General Education*, *58*(3), 129–151. doi:10.1353/jge.0.0047

Trautwein, U., Marsh, H. W., Nagengast, B., Lüdtke, O., Nagy, G., & Jonkmann, K. (2012). Probing for the multiplicative term in modern expectancy–value theory: A latent interaction modeling study. *Journal of Educational Psychology*, *104*(3), 763–777. doi:10.1037/a0027470

Van Barneveld, C., Pharand, S.-L., Ruberto, L., & Haggarty, D. (2013). Student motivation in large-scale assessments. In M. Simon, K. Ercikan, & M. Rousseau (Eds.), *Improving large-scale assessment in education: theory, issues and practice* (pp. 43–61). New York, NY: Routledge.

Vollmeyer, R., & Rheinberg, F. (2006). Motivational effects on self-regulated learning with different tasks. *Educational Psychology Review*, *18*(3), 239–253. doi:10.1007/s10648-006-9017-0

Weirich, S., Penk, C., Hecht, M., Roppelt, A., & Böhme, K. (2015). Item position effects are moderated by changes in test-taking effort. *Manuscript submitted for publication*.

Wigfield, A., & Cambria, J. (2010). Students' achievement values, goal orientations, and interest: Definitions, development, and relations to achievement outcomes. *Developmental Review*, *30*(1), 1–35. doi:10.1016/j.dr.2009.12.001

Wigfield, A., & Eccles, J. S. (1992). The development of achievement task values: A theoretical analysis. *Developmental Review*, *12*(3), 265–310. doi:10.1016/0273-2297(92)90011-P

Wigfield, A., & Eccles, J. S. (2000). Expectancy–value theory of achievement motivation. *Contemporary Educational Psychology*, *25*(1), 68–81. doi:10.1006/ceps.1999.1015

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In A. Wigfield & J. S. Eccles (Eds.), *Development of achievement motivation* (pp. 91–120). San Diego: Academic Press.

Wise, S. L. (2006). An investigation of the differential effort received by items on a low-stakes computer-based test. *Applied Measurement in Education*, *19*(2), 95–114. doi:10.1207/s15324818ame1902_2

Wise, S. L. (2009). Strategies for managing the problem of unmotivated examinees in low-stakes testing programs. *The Journal of General Education*, *58*(3), 152–166. doi:10.1353/jge.0.0042

Wise, S. L., Bhola, D. S., & Yang, S. (2006). *Taking the time to improve the validity of low-stakes tests: The effort-monitoring CBT*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Wise, S. L., & DeMars, C. E. (2005). Low examinee effort in low-stakes assessment: Problems and potential solutions. *Educational Assessment*, *10*(1), 1–17. doi:10.1207/s15326977ea1001_1

Wise, S. L., & DeMars, C. E. (2006). An application of item response time: The effort-moderated IRT model. *Journal of Educational Measurement*, *43*(1), 19–38. doi:10.1111/j.1745-3984.2006.00002.x

Wise, S. L., & Kong, X. J. (2005). Response time effort: A new measure of examinee motivation in computer-based tests. *Applied Measurement in Education*, *18*(2), 163–183. doi:10.1207/s15324818ame1802_2

Wise, S. L., Pastor, D. A., & Kong, X. J. (2009). Correlates of rapid-guessing behavior in low-stakes testing: Implications for test development and measurement practice. *Applied Measurement in Education*, *22*(2), 185–205. doi:10.1080/08957340902754650

Wise, S. L., & Smith, L. F. (2011). A model of examinee test-taking effort. In J. A. Bovaird, K. F. Geisinger, & C. W. Buckendahl (Eds.), *High-stakes testing in education: science and practice in K-12 settings* (1st ed., pp. 139–153). Washington, DC: American Psychological Association.

Wise, V. L. (2004). *The effects of the promise of test feedback on examinee performance and motivation under low-stakes testing conditions* (Unpublished doctoral dissertation). University of Nebraska–Lincoln, Lincoln.

Wise, V. L., Wise, S. L., & Bhola, D. S. (2006). The generalizability of motivation filtering in improving test score validity. *Educational Assessment*, *11*(1), 65–83. doi:10.1207/s15326977ea1101_3

Wolf, L. F., & Smith, J. K. (1995). The consequence of consequence: Motivation, anxiety, and test performance. *Applied Measurement in Education*, *8*(3), 227–242. doi:10.1207/s15324818ame0803_3

Wolf, L. F., Smith, J. K., & Birnbaum, M. E. (1995). Consequence of performance, test motivation, and mentally taxing items. *Applied Measurement in Education*, *8*(4), 341–351. doi:10.1207/s15324818ame0804_4

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Eds.), *Educational measurement* (4th ed., pp. 111–153). Westport, CT: Praeger Publishers.

Zilberberg, A., Finney, S. J., Marsh, K. R., & Anderson, R. D. (2014). The role of students' attitudes and test-taking motivation on the validity of college institutional accountability tests: A path analytic model. *International Journal of Testing*, *14*(4), 360–384. doi:10.1080/15305058.2014.928301